Response Times and Decision-Making

Chris Donkin[1] and Scott D. Brown[2]
[1] School of Psychology, University of New South Wales
[2] School of Psychology, University of Newcastle

Response times have been very informative for the understanding of mental processes, for many years. The most useful analyses of response times have been those based on cognitive theories of decision-making, known as evidence accumulation models. We review the history of decision-making models, and the empirical phenomena which have guided their development. We focus particularly on the common elements of the models, as they represent theoretical agreement about the most fundamental elements of decision-making theory. We also review the practical usage of evidence accumulation models as data-analysis tools, with a discussion of the strengths and weaknesses of this popular approach. While popular, model-based analysis of response time data can be challenging, and so we review developments which make those analyses easier, and also discuss common approaches to the most common problems in plotting, parameter estimation, and model selection.

Introduction

Much of experimental psychology uses accuracy and response time (RT) data to make inferences about the processes underlying performance. These data are used in many forms, from simple mean RT or accuracy on their own, through to complete joint distributions over both measures. RT data are used in many different research paradigms, including classic areas of basic and applied psychology such as: memory; executive function (inhibitory control, and task switching); reading; numeracy; categorisation; reasoning; intelligence research; attention; visual and auditory perception; animal cognition; clinical psychology; and human factors. The classic textbook on response times, by Luce (1986), reviews these topics.

Standard paradigms for investigating the above kinds of decisions involve "speeded choice". Participants are repeatedly asked to make simple decisions with a focus on both the accuracy and speed of their responses. For example, participants might be asked to decide whether the number of items in a simple display is more or less than some criterion value (Ratcliff & Rouder, 1998). The data from speeded choice paradigms include both RT and accuracy, and it is important that those two variables be considered jointly. A central reason for this is the potential trade-off between how long a response takes to make and the likelihood that the response will be correct. The long-studied speed-accuracy tradeoff describes how responses made quickly are more likely to be incorrect (Wickelgren, 1977; Schouten & Bekker, 1967; Pachella, 1974), making independent analyses of each dependent variable problematic. For example, imagine a comparison in which people from Group A were able to respond, on average, in 500ms, and Group B in 1,000ms. It is tempting to infer that people from Group A perform better than Group B. What if, however, we subsequently found out that those in Group A made more errors (15% incorrect responses) than Group B (5% incorrect responses). Because Group A were faster but made more errors than Group B it is possible that both groups performed the task equivalently well, but that Group B was more cautious. It is possible that if people in Group A were encouraged to be more cautious, such that they too made errors only 5% of the time, that their mean RT might also be 1000ms.

In this simple example, the speed-accuracy tradeoff was easy to spot, but it is not always so. Frequently, there can be very large differences in mean RT which occur with very small – even statistically nonsignificant – differences in accuracy. The standard approach of submitting accuracy and mean RT to separate statistical tests does not always address the problem. Even in the simplest cases, the standard approach provides no guidance on the central question of interest: how to combine RT and accuracy to judge the overall performance level. The above examples example demonstrate that there are many determinants of task performance, beyond just one's basic ability to perform the task, such as caution, bias, or even the time to make the required motor response.

The first key step in understanding the underlying causes of differences in RT and accuracy comes from analyzing not just mean RT, but the joint distribution over RT and accuracy. This joint distribution specifies the probability of making each response (say, correct vs. incorrect responses, or "bright" vs. "dark" responses) at all different RTs. The second key step is to interpret these joint distributions by fitting quantitative models of speeded decision-making. There are many quantitative cognitive models which explain RT and accuracy distributions in terms of latent variables representing decision-making processes. The most successful models of RT and accuracy ("choice RT models") are the evidence accumulation (or sequential sampling) models, including: the diffusion model, (Ratcliff, 1978); the EZ diffusion model (Wagenmakers, van der Maas, & Grasman, 2007); the Poisson accumulator model (Pike, 1966; P. L. Smith & Vickers, 1988; Van Zandt, Colonius, & Proctor, 2000); the leaky competing accumulator model (Usher & McClelland, 2001); the Ising decision model (Verdonck & Tuerlinckx, 2014); the urgency gating model (Cisek, Puskas, & El-Murr, 2009); and the ballistic accumulator models (Carpenter & Reddi, 2001; Brown & Heathcote, 2005, 2008).

All evidence accumulation models share the basic assumption that participants sample information from the environment. This information is then taken as evidence for one of the competing responses. Evidence is accumulated until it reaches some threshold level for one of the potential responses. That response is then chosen, with the time taken for evidence to reach the threshold being the decision time component of the RT (Stone, 1960). To explain the variability in RTs and in response choices (i.e., errors), the models assume that evidence accumulation is noisy. This noise means that on some trials evidence for incorrect responses will reach threshold before evidence for the correct response.

Decision-making models make predictions for the joint distribution over RT and choice, and these predictions are defined by latent parameters which represent processes underlying how decisions are made. Of these variables, three are common across all variants of evidence accumulation models and are often of central research interest (Wagenmakers et al., 2007). The three variables are rate of processing, response caution and non-decision time. Rate of processing, often called drift rate, refers to the speed at which evidence for a response is accumulated, and is a measure of how well the task is being performed. Response caution refers to how much evidence is required before a response is made, and is most often responsible for producing a trade-off between the speed and accuracy of responses. By setting a large threshold for how much evidence is required before making a response, a participant will wait longer to make a decision. Waiting this extra time means that the response is more likely to be correct, as noise in the evidence accumulation process will be integrated out with time. When the threshold is set low, however, responses will be faster but more vulnerable to noise in the system, and hence more likely to be incorrect. Non-decision time refers to the time taken for all components of RT which are not part of the evidence accumulation process. The non-decision time is added to the decision time produced by the evidence accumulation process to give a predicted RT, on the basis of a strictly-serial assumption. Non-decision time is most often represented as a simple additive constant value, although some models assume that uniform noise is added (Ratcliff & Tuerlinckx, 2002; Verdonck & Tuerlinckx, 2016).

Though all evidence accumulation models have some form of these three latent variables, their exact form within any particular model varies substantially. The different choice RT models also make considerably different assumptions about what noise is necessary to account for RT and accuracy data. What follows is an overview of some of the more popular choice RT models, with particular focus on two things: how the three aforementioned latent variables are implemented, and which sources of noise are assumed to be important enough to model.

Overview of Decision-Making Models

There have been dozens of different evidence accumulation models developed and tested against data, ranging from very simple random walks (Stone, 1960) through to detailed descriptions of complex neural circuits (Lo & Wang, 2006; Frank & Claus, 2006; Frank, 2006). We have organized our brief review of some of these models into two sections, according to whether the models posit multiple, racing, accumulators, or a single accumulator between multiple boundaries. To help keep track of the relationships between these models, Figure 1 provides a schematic illustration of the relationships between some of the models. This figure is similar to Figure 1 of Ratcliff and Smith (2004) and to Figure 4 of Bogacz, Brown, Moehlis, Holmes, and Cohen (2006), both of which the reader might find useful for more detailed taxonomies of some parts of the model space.

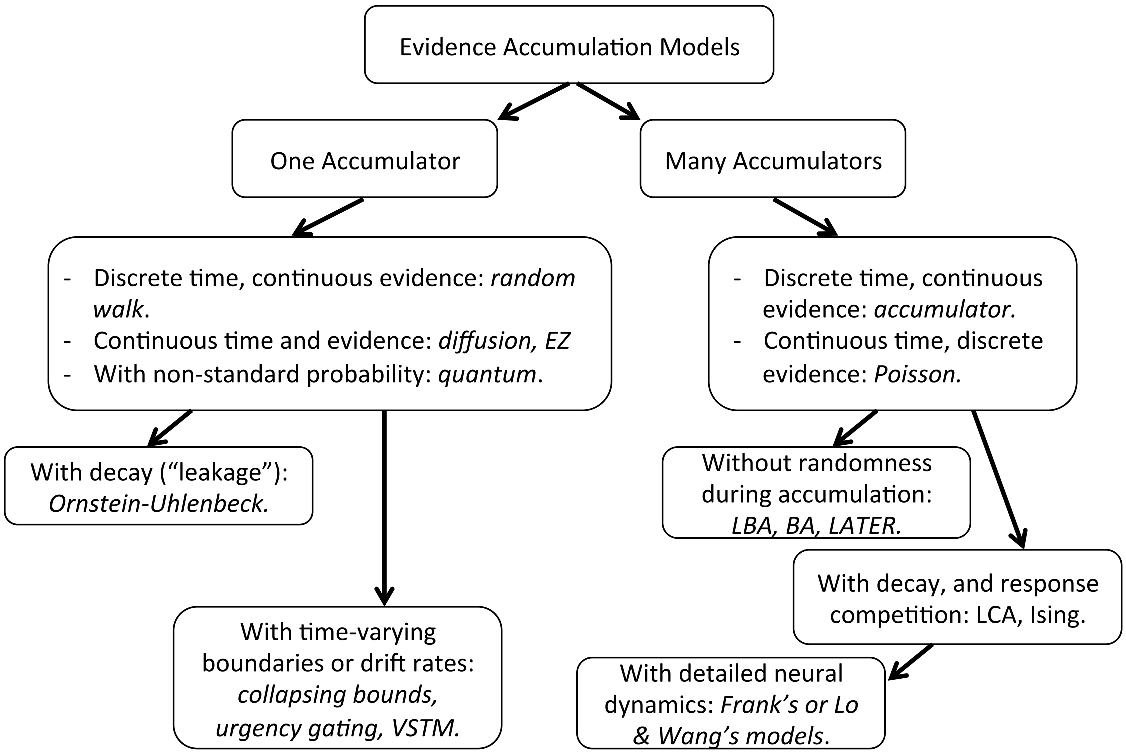

Figure 1. Schematic illustration of the relationships between some evidence accumulation models. Mostly, the complexity of the models increases from top to bottom of the figure.

Single Accumulator Models. One of the first attempts to model RT distributions was the random walk model (Stone, 1960; Link & Heath, 1975; Laming, 1968; Bogacz et al., 2006). In a random walk process, time passes in discrete time steps of length Δt. During each time step some evidence is extracted from the environment suggesting which of the two possible responses (say, A or B) is correct. This evidence then increments a counter, say x, such that if the evidence supports response A the value of x increases, and if the evidence supports response B then x decreases. When x equals some threshold value, say a for response A and 0 for response B, then that particular response is made, and the number of time intervals of size Δt determines the time taken for the decision to be made.

Evidence accumulation begins at some intermediate value, $0 \leq z \leq a$. If there is no bias towards either responding A or B then $z = \frac{a}{2}$, the midpoint between the two response threshold values. If there is bias towards one particular response then evidence accumulation will start closer to that response threshold value. During each time step the amount of evidence added to or subtracted from x is sampled from a normal distribution with mean δ and standard deviation s. This δ value is the drift rate parameter in a random walk model because it indicates the average rate at which evidence accumulates towards boundary a or 0. A positive drift rate indicates more evidence for response A, while a negative drift rate suggests more evidence for response B. Drift rates closer to zero lead to slower and more error-prone responses because the accumulation process is influenced predominantly by the variability in drift rate between time steps. The standard deviation of the drift increments is frequently fixed at either $s = 1$ or $s = 0.1$, to constrain a scaling property of the model (Donkin, Brown, & Heathcote, 2009b).

The size of a reflects response caution in the random walk model. If a is small, then two response boundaries are close together. This means that little evidence is required to trigger a response, and errors due to the stochastic nature of evidence accumulation will occur often. On the other hand, if a is large, then fewer errors will be made, but the accumulation process will take longer to reach a threshold, and so responses will be slower. Non-decision time in the random walk model, T_{er}, is added to the decision time to give the standard RT.

In order to account for performance in recognition memory tasks,Ratcliff (1978) studied a continuous time version of the random walk model. This model (see also Feller, 1971) assumed continuous evidence accumulation by investigating the limit of small time steps, and small drift rates: $\Delta t, \Delta d \rightarrow 0$ (see BOX: HOW THE DIFFUSION MODEL WORKS for more detail). The accumulation of evidence in the continuous version of a random walk model is also referred to as a Wiener process, or Brownian motion, or a diffusion model. Ratcliff also made a very important addition to the basic model: to accommodate the empirical finding that the mean RT for error responses is often slower than the mean RT for correct responses in recognition memory experiments, Ratcliff added the additional assumption that drift rate δ varied from trial-to-trial according to a normal distribution with mean v and standard deviation η. This assumption allowed the model to account for slow error responses, via a mixture argument: correct responses arise more frequently from large samples of δ, which are also fast, while, incorrect responses arise most frequently from small samples of δ, which are also slow.

Later experiments also showed that error responses from the one experiment could be both faster *and* slower than correct responses when the decisions were high and low in accuracy, respectively (P. L. Smith & Vickers, 1988; Ratcliff, Van Zandt, & McKoon, 1999; Ratcliff & Rouder, 1998). To accommodate this pattern, Ratcliff and Rouder borrowed inspiration from the model of Laming (1968), and added trial-to-trial variability in the starting point of evidence accumulation. Ratcliff and Rouder showed that a diffusion model could predict fast errors if start-point (z) was allowed to vary according to a uniform distribution with mean z and range s_z. Having *both* trial-to-trial variability in start point and drift rate allows a diffusion process to produce both faster and slower error RTs for easy and hard conditions, even within a single block of experimental trials.

To explain changes across experimental conditions in the speed of the very fastest responses, a third source of trial-to-trial variability was later added to the diffusion model.

Ratcliff and Tuerlinckx (2002) added variability in non-decision time. Without this assumption, the diffusion model predicts that, regardless of drift rate, the fastest responses made by participants all take a similar amount of time. This property is sometimes called a "flat leading edge" of the RT distribution, and it is very often observed in data, but is not quite universal. Ratcliff and Tuerlinckx demonstrated that the diffusion model gave better account of empirical data when non-decision time was allowed to vary according to a uniform distribution with mean T_{er} and range s_t. Allowing non-decision time to vary across trials also helped the diffusion model account for performance in the lexical decision task, where relatively large changes in the leading edge were observed across stimulus-based conditions (Ratcliff, Gomez, & McKoon, 2004; Wagenmakers, Ratcliff, Gomez, & McKoon, 2008).

A diffusion model with these three sources of trial-to-trial variability is now the most successful and widely-used model of decision-making, and is due largely to the work of Ratcliff and colleagues (in recognition, this particular implementation of the diffusion model is usually called "the Ratcliff diffusion model"). For reviews of applications of the diffusion model, and also open questions about its ongoing development, see B. Forstmann, Ratcliff, and Wagenmakers (2016); Ratcliff, Smith, Brown, and McKoon (2016).

Apart from the Ratcliff diffusion model, there are alternative diffusion models, such as the Ornstein-Uhlenbeck model (OU: Busemeyer & Townsend, 1992, 1993). The OU process differs from the standard Wiener diffusion model because the evidence total, x, decays back towards a resting value, and away from response thresholds. Ratcliff and Smith (2004) showed that the OU model did not perform as well as the standard Wiener diffusion model in some data sets. Still others have investigated random walk models with non-standard probability calculus, most notably the "quantum random walk" (Busemeyer, Wang, & Townsend, 2006). This approach has the benefit of naturally explaining certain phenomena in which people diverge from standard probability, such as via sequential effects, and in consumer choices.

Wagenmakers et al. (2007) provided simple methods for estimating rate of processing, response caution and non-decision time parameters for a basic Wiener diffusion model (i.e., one that contains none of the three sources of between-trial variability). This method, called the "EZ-diffusion" model, involves the estimation of the a, δ and T_{er} parameters via method of moments, using the mean and variance of RT and the percentage of correct responses. The EZ-diffusion model provides an excellent alternative for users who do not want, or need, the complexity and estimation difficulty of the full Ratcliff diffusion model. Even though the EZ-diffusion model has obvious shortcomings as a theory of decision-making (e.g., it cannot accommodate fast or slow errors), in many situations the EZ-diffusion provides a good account of data, and reliable parameter estimation.

Multiple Accumulator Models. Both random walk and diffusion models are examples of single accumulator models, as evidence is tracked by a single accumulator variable. In contrast, multiple accumulator models use an accumulator for each possible response. The recruitment model (LaBerge, 1962, 1994) was one of the first to use a separate accumulator for each possible response. In the recruitment model, time passes in discrete steps, and on each step a unit of evidence is placed in just one of the available accumulators. Thus, in LaBerges recruitment model both time steps and the increment in evidence are discrete. With this exceedingly constrained structure, the recruitment model failed to account for the shapes of empirical RT distributions for correct and error responses, particularly for conditions in which responses are slow. Vickers and Smith's accumulator model (P. L. Smith & Vickers, 1988; Vickers, 1979) also assumed discrete, equally-spaced time periods, but assumed that the amount of evidence incremented between these time periods is sampled from a continuous distribution (see also the PAGAN model: Vickers & Lee, 2000a, 1998). Conversely, the Poisson counter model (Pike, 1966, 1973; LaBerge, 1994; Van Zandt et al., 2000; Townsend & Ashby, 1983) assumes that the amount of evidence accumulated on each step is fixed but that the time intervals in which evidence arrives varies randomly from step-to-step. We now turn to a more detailed discussion of some of these models.

In the accumulator model of P. L. Smith and Vickers (1988), evidence is accumulated at equally-spaced time steps. At each time step, the amount of evidence to accumulate is sampled from a normal distribution. This evidence value is then compared to a criterion value, and if the evidence is larger than the criterion then the difference between the criterion and the evidence value is added to counter B, and if the evidence is smaller than the criterion then counter A is increased by the same difference. When the evidence in either counter reaches a response threshold then that response is made, and the time taken to make the response is the number of time steps multiplied by a constant which converts time steps to seconds. The distance of the mean of the normal distribution of evidence values from the criterion value is equivalent to the drift rate in the diffusion model, in that it reflects the average amount of evidence accumulated per time. Smith and Vickers showed that an accumulator model with three sources of between-trial variability provided a good account of empirical data. Firstly, the mean of the evidence accrual distribution was assumed to vary from trial-to-trial according to a normal distribution. Secondly, non-decision time was assumed to vary across trials. Thirdly, the response threshold was allowed to vary from trial-to-trial according to an exponential distribution. These three sources of variability correspond closely to the three sources of between-trial variability in Ratcliff's diffusion model.

In the Poisson counter model (LaBerge, 1994; Merkle, Smithson, & Verkuilen, 2011; Pike, 1973; Van Zandt et al., 2000) it is assumed that equal amounts of evidence arrive on each time step, but that the time steps vary in size. The time between when evidence arrives in each accumulator is assumed to be exponentially distributed with separate rate parameters for each possible response. Because the time between evidence arrival is exponential, the rate at which evidence increases in each accumulator is distributed according to a Poisson process. The evidence accumulation process continues until evidence in one of the accumulators reaches a response threshold. Three sources of between-trial variability have been added to the Poisson counter model: in non-decision time; in the rate of arrival of information for each counter; and in the response thresholds. Despite the addition of these sources of variability, the Poisson counter model is unable to produce both fast and slow errors within experimental blocks (Ratcliff & Smith, 2004; Van Zandt et al., 2000).

Usher and McClelland (2001) developed their "leaky competing accumulator" (LCA) in part to address the shortcomings of previous multiple-accumulator models, and partly also to integrate findings about the neuroscience of decision-making with cognitive modeling. The LCA model assumes separate accumulators for each choice response, like other multiple-accumulator models, but also allows evidence in favor of one response to "count against" evidence in favor of other responses, like in the single-accumulator models. The LCA operationalizes this assumption by adding lateral inhibitory connections to an OU model. These connections mean that evidence in one accumulator inhibits the rate of evidence accrual in the other accumulator(s), at a rate proportional to the current amount of evidence in the inhibiting accumulator. In an opposing force, the LCA model also assumes that accumulators "self excite" – that is, a tendency to grow in activation at a rate proportional to current activation. The LCA does not require trial-to-trial variability in drift rate to predict slow error RTs, because of the lateral inhibitory assumption. The LCA was also able to predict fast error RTs in the same way as other models, by assuming that the start point of evidence accumulation in each accumulator varies randomly from trial-to-trial.

Brown and Heathcote (2005) showed that a simplified version of the leaky competing accumulator model, the ballistic accumulator (BA) model, was able to account for all benchmark choice RT phenomena the shape of RT distributions, the speed-accuracy trade-off, as well as both fast and slow errors. The only difference between the BA and Usher and McClelland's (2001) LCA model is that there is no moment-to-moment variability in the evidence accumulation process. In other words, evidence from the environment was not assumed to follow a Wiener or OU process, but was assumed to be noiseless ("ballistic", although those authors should probably have chosen a better word). With between-trial variability in drift rate, and in the start point of evidence accumulation, passive decay and self-excitation of accumulated evidence, and lateral inhibition between accumulators that the BA model was able to predict all the regular benchmark phenomena, and also accommodate empirical data from a simple discrimination task.

Taking this simplification further, Brown and Heathcote (2008) developed the linear ballistic accumulator model (see BOX: HOW THE LBA MODEL WORKS for more details). In the LBA, accumulation was assumed to be free of leakage, excitation and lateral inhibition. All that remained in the model was deterministic linear evidence accumulation, with two sources of trial-to-trial variability: in drift rate and in start points. Quite surprisingly, the LBA was capable of accounting for the shape of RT distributions, the speed-accuracy trade-off, as well as the relative speed of errors. The mathematical simplicity of the LBA means that it is easy to apply to data, and amenable to advanced statistical approaches.

A modern multiple-accumulator model is the Ising decision-maker developed by Verdonck and Tuerlinckx (2014). This theory is based on neurally inspired ideas similar to other competitive accumulator models, such as the LCA (Usher & McClelland, 2001). The Ising decision maker begins with the assumption that there are two pools of neurons representing two different decision options, and that these pool compete in a winner-takes-all fashion. The Ising decision-maker distills many of the important attributes of detailed, neurally plausible models of decision-making (such as that described byLo & Wang, 2006) into a simpler form. A key property of the Ising decision-maker is that these neurons are reduced to a impoverished representation as simply binary on/off elements. This reduction allows for a tractable analysis of the entire competing system, which has not been possible for other neurally-inspired accumulator models.

Interim Summary

The above was a brief and selective summary of decades of work in the development of RT models. Below, the discussion is continued, divided into two sections: Theory and Measurement. In the first section, we focus on RT models as a route to understanding the way in which humans make decisions. We begin by summarizing the core empirical data patterns that have helped discriminate between RT models to date. We then review recent approaches to testing RT models , and discuss some novel extensions to RT models. We finish this section with an overview of the connections between RT models and neuroscience. In the second section, we discuss the use of RT models as a measurement tool. In recent years, RT models have been used increasingly often to measure the latent variables assumed to underlie decision-making, including ability, caution, bias, and non-decision processes. In this section, we discuss the issues associated with using such relatively complex models as measurement models.

Response Time Models as Theory Development

Certain empirical phenomena have proven particularly important in directing the development of RT models as explanations of the cognitive processes which underpin decision-making. These phenomena have helped to narrow down the field of plausible theoretical explanations, and also provided evidence in favor of particular model elements across a wide variety of different theories.

Speed-Accuracy Tradeoff

Except for the ballistic theories, RT models account for the SAT because increased accumulation time allows the effects of within-trial variability in information accumulation to be integrated out. The simplest models, such as the EZ-diffusion and other early versions of the diffusion and random walk models, have only one source of variability – within-trial variability in evidence accumulation. Since this source can be integrated out by raising the decision threshold, those models predict perfect asymptotic accuracy for all decisions. That is, a decision-maker could achieve any desired accuracy by simply making sufficiently slow decisions. However, less than perfect accuracy is almost always observed in practice, even with unlimited decision time. At least two suggestions have been made to allow stochastic models to account for less than perfect asymptotic accuracy. Usher and McClelland (2001) proposed that accumulation is "leaky" so that information is lost during accumulation, and hence accuracy is imperfect (although asymptotic accuracy in information-controlled paradigms can still be infinite (Busemeyer & Townsend, 1992). Ratcliff (1978) added between-trial variability in the input to the diffusion model, thus predicting imperfect asymptotic accuracy. That is, on some trials, the stimulus will be erroneously encoded as favoring the wrong response, and integrating out the within-trial noise will not redress the problem on those trials.

The ballistic models (Brown & Heathcote, 2008, 2005) produce a speed-accuracy tradeoff via a different mechanism. In those models, where there is no within-trial variability in evidence accumulation, extra integration time instead allows the input to overcome noise in the starting points. To illustrate, consider the example LBA model accumulation trajectories in Figure 5. The unit with a smaller input (dashed line) started with larger activation, but with extra integration time, it was overtaken by the unit with a larger input. If the response criterion (horizontal line) were set very low, the model would make the wrong response, because the accumulator corresponding to the wrong response begins with a slight advantage and would reach a low response criterion first. Raising the response criterion (to the value shown) allows sufficient integration time for the accumulator corresponding to the correct response to overcome its initial disadvantage. Extending integration time indefinitely allows all effects of start point variability to be removed. However, even then, asymptotic accuracy is still imperfect because of variability in input strength.

Fast and Slow Errors

The addition of variability in drift rates fixes another problem for the earliest diffusion models, which included only Gaussian accumulation noise: they predicted equal correct and error RT distributions. Equal correct and error RTs are occasionally observed but typically, when response accuracy is emphasized and the decision to be made is relatively difficult, error RTs are slower than correct RTs, a phenomenon we will call "slow errors". The addition of between trial variability in drift rate allows the diffusion model to produce slow errors (Ratcliff, 1978). In contrast, the LCA model of Usher and McClelland (2001) can produce equal correct and error RTs or slow errors, even though it does not include between-trial variability in parameters. The LCA model makes these predictions due to the inclusion of lateral inhibition and leakage.

When simple decisions are required, and response speed is emphasized, an opposite pattern occurs: error RTs are typically faster than correct RTs, called "fast errors" (e.g., Ratcliff & Rouder, 1998; Ratcliff et al., 1999; see Luce, 1986, p.233 for a review). Fast errors require a third source of variability to be incorporated into the diffusion model, between-trial variability in either the criterion or start point (given reasonable constraints on the variability distributions, these changes are identical when integration is linear, as in the diffusion). Start point variability was originally suggested by Laming (1968) as being caused by pre-stimulus accumulation. Usher and McClelland (2001) also incorporated between-trial start point variability into their model in order to account for fast errors, although they did not fit this version of their model to data from an information controlled task, as only slow errors were observed in those data.

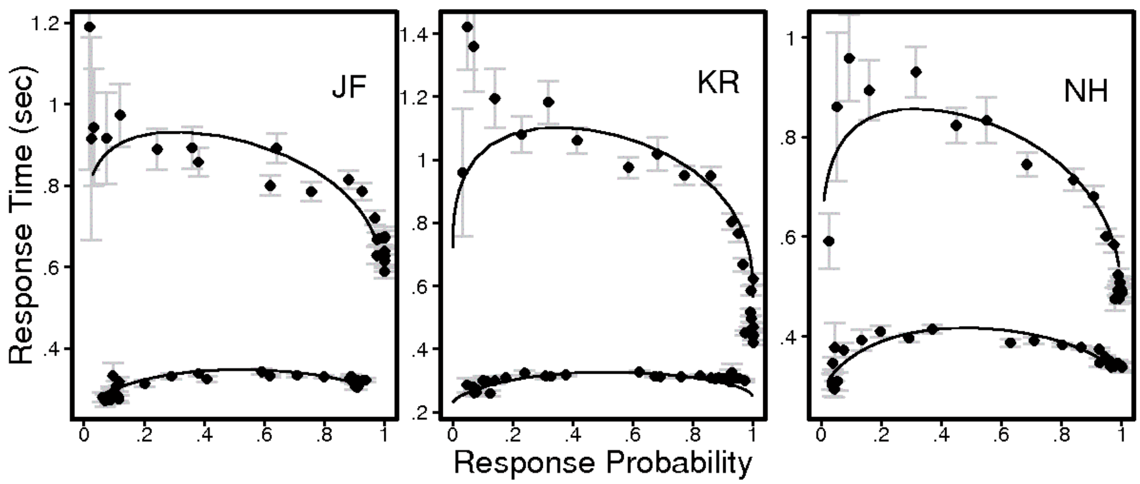

Figure 2. Mean RT (symbols) and predicted mean RT from the LBA model (lines) for three subjects from Ratcliff and Rouder's (1998) experiment. The upper and lower lines are for accuracy and speed emphasis conditions, respectively. Within each condition, there are 33 separate points – one for each level of stimulus brightness. The right side of each plot represents correct responses to very easy-to-classify stimuli, and the left side of each plot represents (very rare) incorrect responses to the same stimuli. The center of each plot shows data from difficult stimuli, which were nearly equally often classified correctly and incorrectly. Bars indicate standard error.

A pattern that has proven particularly diagnostic for selecting models of choice RT is a crossover effect, in which faster and slower error RTs are observed in easy and hard stimulus discrimination conditions respectively, even when these conditions are randomly intermixed from trial to trial. Hence, general choice RT models must be able to accommodate crossovers by changing only stimulus-driven parameters, and not parameters which require strategic control from the decision-maker. Figure 2 illustrates the crossover pattern observed by Ratcliff and Rouder (1998), using a plotting style which has become important in RT research, called a "latency-probability" plot (LP plot: Audley & Pike, 1965). Latency probability plots show mean RT as a function of the probability of a response. Points on the left of the graph represent the lower probability (error) responses and complementary points on the right of the graph represent the higher probability (correct) responses from the same experimental conditions. Sometimes, LP plots are expanded to show more than just the mean RT, by plotting several quantiles of the RT distributions – these are called "quantile-probability", or QP, plots.

The “crossover” pattern in the speed of correct and incorrect choices is evident in Figure 2 in several ways. Data from the accuracy-emphasis condition (upper symbols in each plot reveal uniformly slow errors: each data point on the left side of the graph, representing correct response mean RT for some probability $p > .5$ is a little faster than the corresponding speed for incorrect responses, plotted at $1-p$. The data from the speed-emphasis condition for subject JF (left panel, lower data) show uniformly fast errors: points plotted at probability $p > .5$ are always a bit slower than the corresponding errors plotted at $1 - p$. The speed-emphasis data from subject NH shows a crossover pattern. For every easy decisions, the correct responses (plotted near $p = 1$) are slower than their corresponding error responses (plotted near $p = 0$). For difficult decisions, plotted near the middle of the graph, incorrect responses (such as those at $p = .4$) are slower than the corresponding correct responses (which are plotted at $p = .6$). Most modern RT models are able to accommodate this pattern, by including between-trial variability in various model parameters.

Choices between more than two options.

The vast majority of response-time and decision-making studies have used binary decision tasks, for example “target vs. distractor”, “bright vs. dark”, “many vs. few”, “left vs. right”, and so on. Nevertheless, there are a substantial number of studies that have investigated decisions between more than two response options, and these experiments have yielded their own set of important empirical phenomena. The single most important empirical result from multiple-choice experiments is Hick’s Law (Hick, 1952; Hyman, 1953), which describes how decisions become steadily slower with response alternatives. Hick’s Law can be expressed in a number of ways, but the most simple is that the mean time taken to select a response (i.e., $\overline{RT}$) and the logarithm of the number of choice alternatives (K) are linearly related:

$$\overline{RT} = a + b\log_2(K). \tag{1}$$

Hick’s Law describes data from a wide range of paradigms including speeded perceptual judgments (e.g., Leite & Ratcliff, 2010), eye saccades (e.g., anti-saccades in Kveraga, Boucher, & Hughes, 2002; K.-M. Lee, Keller, & Heinen, 2005), absolute identification (e.g., Lacouture & Marley, 1995; Pachella & Fisher, 1972), manipulations of stimulus-response compatibility (e.g., Brainard, Irby, Fitts, & Alluisi, 1962; Dassonville, Lewis, Foster, & Ashe, 1999), and has even been observed in monkeys (Laursen, 1977) and pigeons (Vickrey & Neuringer, 2000; for additional examples in other paradigms see Brown, Steyvers, & Wagenmakers, 2009; Teichner & Krebs, 1974; ten Hoopen, Akerboom, & Raaymakers, 1982).

Hick’s Law has important implications for theories of decision-making and RT. The single-accumulator models of decision-making, such as the random walk and diffusion models, are naturally restricted to making predictions about only binary choices. In contrast, multiple-accumulator models naturally extend to multiple choice tasks: for a choice between N different responses, the standard assumption is to have N racing accumulators. However, more complex arrangements are possible, for example with accumulators in pairs making pairwise comparisons between different response options. The most pressing difficulty with the standard account is that it fails to predict Hick's Law. All else being equal, if more response options are added, then more accumulators race to the threshold, and so the probability that one of them will finish very quickly becomes larger. This effect is called "statistical facilitation", and predicts the opposite of Hick's Law, faster RT with more choices.

Many different ideas have been proposed to address this shortcoming. Usher, Olami, and McClelland (2002) proposed that RTs slowed in larger choice sets simply because decision-makers became more cautious, and lifted their response thresholds. Hawkins, Brown, Steyvers, and Wagenmakers (2012) investigated models based on continuous hypothesis testing of the different response alternatives, which led to naturally slower responses with more choices. Other models have been developed for specific and interesting multiple-choice paradigms, such as absolute identification (Brown, Marley, Donkin, & Heathcote, 2008; Lacouture & Marley, 1995) and confidence ratings (Ratcliff & Starns, 2013, 2009; Pleskac & Busemeyer, 2010). A common assumption in these models is some form of normalization - the total amount of some resource is spread across the different response options, thereby reducing processing speed when more response options are added, and accommodating Hick's Law.

Teodorescu and Usher (2013) made a systematic and thorough investigation of many different ways of instantiating *inhibition*. When different response alternatives inhibit one another, then adding more alternatives creates more inhibition, slower responses, and Hick's Law. Inhibition can be added either at the level of competition between outputs, or inputs, or both. It can be added via normalization, or lateral competition, or other methods. Teodorescu et al. investigated all of these options, and concluded that only a select few of them were able to predict Hick's Law.

One of the challenges faced in research into multiple choice decisions and Hick's Law concerns the decision tasks used. It is not easy to generate a decision task that allows a large number of alternative decisions (say, more than eight) without introducing unwanted elements to the task, such as large memory loads, or perceptual limitations. These problems limit the extent to which data from multiple-choice tasks can be used to draw general conclusions about decision-making; conclusion which apply beyond just the particular task in question. Similar concerns apply to the "extended judgment" task, used by Teodorescu and Usher (2013), Hawkins et al. (2012), Usher and McClelland (2001), and many others since its introduction by Vickers (1979). This task slows down decision-making by presenting a long series of elements, and having the decision-making make a response based on the statistics of the whole sequence. This set-up allows very detailed analysis and powerful model discrimination (Pietsch & Vickers, 1997), but leaves open questions about the generality of the conclusions to more standard decision-making. Teodorescu and Usher (2013) were able to make similarly powerful model discriminations, but also only by assuming very particular mappings between physical stimulus magnitudes and internal psychological magnitudes, and between potential responses and model accumulators.

A different, and probably specialized, kind of choice between more than two options is about decision confidence. A long line of research has investigated the ways in which confidence about a decision is influenced by properties of the decision stimulus, and how the confidence and decision questions are asked. Evidence accumulation models with multiple racing accumulators have a natural way in which confidence might be expressed, sometimes known as the "balance of evidence" hypothesis (Vickers, 1979; Vickers & Lee, 2000b). The balance of evidence hypothesis is that the confidence in a decision is determined by the difference between the amount of evidence in the winning vs. losing accumulators. Difficult decisions will typically lead to the losing accumulator having almost as much accumulated evidence as the wining accumulator, and this small difference will engender low confidence in the decision.

In contrast to the relatively settled notions and broad agreement about the basic components of decision making by evidence accumulation, there is disagreement about the components of confidence judgments. Pleskac and Busemeyer (2010) have developed a modern account of decision confidence based on the balance of evidence hypothesis, and this account fits a wide range of data from decision making and confidence rating experiments. However, Ratcliff and Starns (2013) and Moran, Teodorescu, and Usher (2015) have developed quite different models of confidence that account for many of the same phenomena, and it is not yet clear which of these different approaches is best. While Pleskac and Busemeyer's model hinges on the balance of evidence hypothesis, Ratcliff and Starns treat a confidence rating task as a choice between many alternatives representing different categories of confidence ("low", "medium", ...) and Moran et al. employ collapsing decision boundaries (see next section).

Efforts to distinguish different accounts of confidence have focussed on the identification of qualitative data patterns that might be accommodated by just one of the models, and not the others. These empirical "benchmarks" (or "hurdles") that models of confidence must meet have been growing in number and complexity, and there is not yet a resolution to the debate. The difficulty of the problem has been compounded by the use of different basic empirical paradigms, which seem to favor one account over another. For example, Pleskac and Busemeyer (2010), and others, ask participants to provide a confidence rating directly after making a choice: e.g. a participant might first decide in favor of response "A", and then describe their confidence as "high". In contrast, Ratcliff and Starns (2013) ask participants to make their choice and their confidence judgment simultaneously: e.g. a participant might choose the response option labeled "A: high", as opposed to "B: high", or "A: medium" and so on. Both procedures have advantages, but it is not easy to map data from one paradigm onto theories intended for the other.

Non-stationary decision processes.

All of the RT models reviewed so far are "time homogeneous" – they make the assumption that the rules of evidence accumulation and decision processing do not change as decision time passes. For many decades, such models have provided detailed accounts of decision-making data. More complex time inhomogeneous models have recently been proposed and become especially popular in some neurophysiological studies of primates (e.g., Churchland, Kiani, & Shadlen, 2008; Ditterich, 2006a; Drugowitsch, Moreno-Bote, Churchland, Shadlen, & Pouget, 2012) but not all (e.g., Purcell, Schall, Logan, & Palmeri, 2012). These models are also sometimes known as "non-stationary" or "dynamic" decision models, reflecting that they implement a constantly-changing decision strategy. The most-explored approach is to have the decision boundaries decrease with decision time, which means that the quantity of evidence required to trigger a decision decreases with time. This is often called a "collapsing bounds" assumption.

Diffusion models typically assume fixed decision boundaries; the amount of evidence required to trigger a decision does not change with time (see the response threshold boundaries in Figures 4 and 5). This approach is statistically optimal in that it leads to the fastest mean decision time for any fixed error rate in a single condition, and constant information cost over time (Wald & Wolfowitz, 1948). The collapsing boundaries assumption suggests instead that the diffusion model's boundaries move closer together, or that the LBA model's boundaries move closer to zero as decision time passes (Bowman, Kording, & Gottfried, 2012; Ditterich, 2006a, 2006b; Drugowitsch et al., 2012; Milosavljevic, Malmaud, Huth, Koch, & Rangel, 2010; Thura, Beauregard–Racine, Fradet, & Cisek, 2012). Collapsing boundaries are also statistically optimal under different assumptions about the stimulus environment, the decision-maker's goals and the cost of passing time (Ditterich, 2006a).

While the collapsing boundaries idea is interesting, and has attractive statistical properties regarding optimality, the data mostly speak against this assumption. In the most extensive investigation so far, Hawkins, Forstmann, Wagenmakers, Ratcliff, and Brown (2015) compared models with static versus dynamic response boundaries in a large survey. Overall, data from nine experiments provided strong support for the conventional, fixed bound model. There was evidence in favor of collapsing boundaries or urgency signals for a small proportion of human subjects (mostly from one experiment). Interestingly, there was substantial support for models with collapsing boundaries in studies using monkeys. This result suggests caution in generalizing from non-human primate studies of decision-making to human psychology.

Recently, the basic understanding of decision-making based on evidence accumulation has been challenged by another interesting proposal of non-stationarity, from Cisek et al. (2009) and Thura et al. (2012). This is the "urgency gating model", which goes beyond non-stationarity and drops the central component of the EAMs, by assuming that environmental evidence is *not* accumulated over time. Instead, the UGM tracks novel sensory information, which varies from moment-to-moment, and multiplies this information by an urgency signal that grows with decision time. These multiplied samples are simply monitored until any sample exceeds a decision threshold. The UGM is an original and insightful proposal that has already had important impacts on the field (for similar approaches see Hockley & Murdock, 1987, and accompanying critique from Gronlund & Ratcliff, 1991). Despite the intrinsic interest of the proposal, there are mathematical issues yet to be resolved with the idea of urgency gating (Hawkins, Wagenmakers, Ratcliff, & Brown, 2015). More importantly, the evidence from both human and monkey data seem to support urgency gating models even less than they support collapsing bounds models (Hawkins, Forstmann, et al., 2015).

Response Times in Cognitive Science and Neuroscience

The field of cognitive neuroscience initially sought to map changes in the brain as they related to cognition, using neural measurements obtained through event–related potentials (ERPs; e.g., Sutton, Braren, Zubin, & John, 1965; Hillyard, Hink, Schwent, & Picton, 1973), the magnetoencephalogram (MEG; e.g., Brenner, Williamson, & Kaufman, 1975), functional magnetic resonance imaging (fMRI; e.g., Belliveau et al., 1991), and single–unit recordings in non–human primates (e.g., Hanes & Schall, 1996; Schall, 2001; Shadlen & Newsome, 1996). As progressively more precise measures of the inner workings of the brain became available, researchers have become increasingly capable at understanding the neural determinants of cognitive processes.

Some research paradigms have well–specified and tractable mathematical models of cognition, and also well–developed methods for neural measurement, including decision making. An important change in the development of decision–making models over the past twenty years has been a steady "tightening" of the link between neural and behavioral data (for discussion of linking behavioral and neural data, see Teller, 1984). Early models of simple decision–making linked behavioral and neural data loosely, by constraining the development of behavioral models to respect data from neural measurements. For example, the leaky competing accumulator model developed by Usher and McClelland (2001) was structurally constrained to include components supported by neural investigations, such as lateral inhibition between accumulating units, and passive decay of accumulated evidence. These links were included as part of the model development process, and thereafter there was no further attempt to link neural with behavioral data.

Subsequent models tested the links via qualitative comparisons between predictions for corresponding neural and behavioral data sets. This kind of linking was very common in early research into decision–making with fMRI methods, in which predictions were based on the assumption that an experimental manipulation will influence one particular model component, which leads naturally to predictions for the behavioral data, and also for the neural data (via the hypothesized link). Predictions most frequently take the form "in condition A vs. B, behavioral measure X should increase while neural measure Y decreases". Support for the predictions is taken as evidence in favor of the model, including the hypothesized link. As an example, Ho, Brown, and Serences (2009) tested predictions generated from decision–making models via hypothesized neural links. In one part of their study, Ho et al. manipulated the difficulty of a decision-making task and hypothesized that this should result in a change in the speed of evidence accumulation in an evidence accumulation model. By examination of the model coupled to a standard model for haemodynamic responses, Ho et al. generated predictions for the blood–oxygen–level dependent (BOLD) response profile within regions that are involved in perceptual decision making. These predictions were compared with data from an fMRI experiment, which lent support to some accounts over others.

Linking via the testing of qualitative hypotheses was later surpassed by quantitative approaches, which provided a tighter link between neural and behavioral data. The most common example of quantitative linking in decision–making models takes parameters of the decision–making model, estimated from behavioral data, and compares them against the parameters of a descriptive model estimated from the neural data. For example, B. U. Forstmann et al. (2008) correlated individual subjects' model parameters, estimated from behavioral data, against blood–oxygen–level dependent (BOLD) parameter estimates; subjects with large changes in threshold parameters also showed similarly large changes in BOLD responses.

Most recently, there have been efforts to link neural and behavioral decision–making data even more tightly, by combining both data sets in a single model–based analysis. This approach has culminated in models such as that developed by Purcell et al. (2010) which uses neural measurements as a model input in order to predict both behavioral measurements and a second set of neural measurements. This provides a simultaneous description of neural and behavioral data sets, as well as explicating the links between them. A less detailed, but more general approach was developed by Turner, Forstmann, et al. (2013), in which neural and behavioral models are joined by allowing their parameters to covary. Turner, Forstmann, et al.'s approach is a "joint" model, in the sense that it allows symmetric information flow: behavioral data can influence the neural parameter estimates, and neural data can influence the behavioral parameter estimates.

Examples of Cognitive Neuroscience linked with RT Models. The following is a brief and incomplete review of research that links cognitive models of RT and decision-making with neuroscientific data. The list is organized, approximately, in increasing order of "tightness" in the link between the two data streams. Some of the material is an abridged version of a more complete review, from de Hollander, Forstmann, and Brown (2015).

The leaky competing accumulator model (LCA) of Usher and McClelland (2001) included structural elements such as mutual inhibition between competing accumulators, motivated by neural data which demonstrate the prevalence of inhibitory connections between nearby neurons within the same cortical stratum. Evidence in favor of these links was inferred by the observation that the resulting cognitive model provided a good fit to behavioral data. P. L. Smith (2010) showed that a plausible model of how neurons encode sensory information at very short time scales (a Poisson shot noise process), converges, under reasonable assumptions, to a Ornstein-Uhlenbeck velocity process. The integrated version of this process is, in turn, indistinguishable from a standard diffusion model (Ratcliff, 1978; Ratcliff & McKoon, 2008).

Hanes and Schall (1996) recorded single-cell activity in the frontal eye fields (FEF) in behaving macaques. The activity of "movement neurons" predicted the execution of saccades. Hanes and Schall (1996) showed that the ramping activity of these neurons preceding a saccade always ended with the same firing rate, but the rate of increase of firing rate was variable. Hanes and Schall (1996) interpreted their findings as showing that variability in RT could be explained by variability in drift rate as opposed to variability in threshold of the decision-making process. More and more electrophysiological work has since been interpreted in the framework offered by evidence accumulation models, reviewed by Gold and Shadlen (2001) and B. U. Forstmann et al. (2008).

Links between neural data and evidence accumulation models have also been drawn using fMRI methods. For example, Ho et al. (2009) hypothesized that areas that implement evidence accumulation during a perceptual decision-making task should show delayed and longer activation during difficult trials, compared to easy trials. They identified areas where the shape of the HRF differed substantially between conditions, by testing for interactions between task difficulty and BOLD activity at a set of multiple timepoints throughout the trial. This prediction was supported, at least in averaged data.

An interesting way to link evidence accumulation models of RT with neural data is by relating variability between participants in parameter estimates with variability between those same participants in neuroimaging data. For example, in an fMRI study of decision-making, B. U. Forstmann et al. (2008) instructed subjects to stress either the speed or accuracy of their decisions. The difference in BOLD-activity between accuracy- and speed-stressed trials in the striatum and the pre-supplementary motor area (pre-SMA) was correlated across subjects with the difference in model parameters related to response caution, estimated from behavioral data via the LBA model. In other words, participants who made large changes in their cognitive settings (for speed vs. caution) also showed large changes in fMRI responses, and vice versa. Using a similar across-subjects approach, Mulder, Wagenmakers, Ratcliff, Boekel, and Forstmann (2012) used probabilistic payoffs to shift the decision biases of participants. As usual, these shifts were explained in a perceptual decision making model (the diffusion model) as a shift in the starting point parameter – responses favored by bias were represented as having starting points for evidence accumulation that were closer to the response threshold. Mulder et al. showed that estimates of the start point, taken from behavioral data, were correlated with the difference in fMRI activity between biased and unbiased trials in frontoparietal regions involved in action preparation.

An alternative to the between-subjects approach is to link within-subject variability from neural and behavioral data by splitting the data on a neural measure and fitting a cognitive model to the subsets of behavioral data. Ratcliff, Philiastides, and Sajda

(2009) studied a perceptual decision-making task (houses vs. faces) and identified EEG components that classified trials as hard or as easy. Ratcliff et al. took trials from each single stimulus difficulty condition (in which nominal stimulus difficulty was constant) and applied a median split based on the amplitude of the EEG-component. Even though nominal stimulus difficulty was identical, estimated drift rates were lower in the trials with lower amplitude than trials with a higher EEG amplitude.

Even more recent approaches to linking evidence accumulation models to neural data start with the neural signal, and use this as input to an extended evidence accumulation model. Cavanagh et al. (2011) estimated, separately for each trial in a decision-making experiment, the power in the theta frequency band from recorded EEG signals. These single-trial estimates of theta power were then used to inform parameter estimates in an extended version of the diffusion model (HDDM; Wiecki, Sofer, & Frank, 2013). This model allowed different estimates of the threshold parameter on different trials, and a covariate model to assess the association of single-trial theta power with single-trial threshold estimates.

A similar approach to that of Cavanagh et al. was developed in parallel by Turner, Forstmann, et al. (2013) (see also Turner, van Maanen, & Forstmann, 2014). Also in this "joint modeling approach", neural measures were used in addition to behavioral measures as input to an extended cognitive model. Turner et al.'s approach took the covariate-based analysis further, allowing for a general covariance matrix to link parameters of a behavioral model (the LBA model of decision-making) with the parameters of a neural model (a GLM). This approach supports more exploratory analyses, allowing the identification of different mappings from cognitive parameters to neural measures by studying the covariance matrix of the joint normal distribution; if a cognitive parameter is related to some neural measure, the covariance parameter that links them will be non-zero. Turner, Forstmann, et al. (2013) showed, using the data of B. U. Forstmann et al. (2010), that this approach can find robust correlations between white-matter strength between pre-SMA and striatum, measured by diffusion-weighted magnetic resonance imaging (dMRI).

Response Time Models as Measurement Tools

Most RT models have some parameters that share a common interpretation in terms of the processes that underlie simple decisions: ability, caution, bias, and non-decision processes. These parameters can be used to understand the influence of particular experimental manipulations, real-world interventions, clinical disorders, or other differences of interest. The general approach of using the parameters of quantitative models to describe differences that underlie empirical data has been dubbed "cognitive psychometrics" (J. B. Smith & Batchelder, in press; Batchelder, 1998; Batchelder & Riefer, 1999). RT models have been used extensively for this purpose, with the popularity of this approach increasing.

The typical approach is to run an experiment in which one or more variables are manipulated. This manipulation will have some influence on the joint distribution of RT and accuracy. RT models are then fit to these empirical data, and the differences across experimental conditions are re-interpreted in terms of the model's parameters. This approach relies on being able to estimate the parameters of RT models, and also being able to discern which parameters of the models differ across experimental conditions. We now give a brief overview of existing methods for both issues.

Parameter Estimation

In recent years, with the benefits of cognitive psychometrics becoming more apparent to those outside the field of quantitative psychology, there have been valiant efforts to make the model estimation process more accessible. Some early attempts included written guides and tutorials on fitting RT distributions (Van Zandt, 2000; P. L. Smith, 2000; Ratcliff & Tuerlinckx, 2002). Taking a slightly different approach, Wagenmakers et al. (2007) offered the EZ-diffusion model, and the EZ2 model (Grasman, Wagenmakers, & van der Maas, 2009), as simple ways to estimate parameters for a choice RT model. By working with greatly-simplified RT models, Wagenmakers et al. were able to provide relatively simple formulae that transform mean RT, variance of RT and the proportion of correct responses into estimates of the drift rate, response threshold and non-decision time. The simplified models allowed no between-trial variability (i.e. in drift rate, start point or non-decision time). Such a simplification means that the model no longer gives a full account of benchmark choice RT data. In practice, however, this cost is offset by the fact that researchers in applied areas outside of quantitative psychology benefit greatly from being able to model their data using relatively simple calculations which require no iterated fitting.

Around the same time as the EZ-diffusion model became available, software which made it easier to use the full Ratcliff diffusion model also began to appear: DMAT, (Vandekerckhove & Tuerlinckx, 2008), and fast-DM (Voss & Voss, 2007, 2008). The latest iterations of these packages offer a full range of frequentist methods for estimation including maximum-likelihood, χ^2, and Kolmogorov-Smirnov methods. While maximum-likelihood methods are most efficient, in theory, RT models are particularly susceptible to fast outliers (i.e., responses quicker than those yielded by the true decision-making process). As such, the χ^2 and Kolmogorov-Smirnov methods tend to be more popular.

Recent years have seen the rise of Bayesian methods for parameter estimation (M. D. Lee & Wagenmakers, 2014) for cognitive models. Vandekerckhove, Tuerlinckx, and Lee (2011) give an overview of hierarchical Bayesian estimation for the Ratcliff diffusion model. Bayesian approaches have a clear advantage over frequentist approaches in that they give the full distribution of likely parameter values, in addition to allowing one to incorporate prior information about parameter values (e.g., Matzke & Wagenmakers, 2009). Furthermore, Bayesian methods make it easier to investigate hierarchical extensions of the model, wherein the estimation of an individual's parameters is informed by the estimates of the other participants in the experiment. Wiecki et al. (2013), Wabersich and Vandekerckhove (2014), (Turner, Sederberg, Brown, & Steyvers, 2013) and Donkin, Brown, and Heathcote (2009a) have provided code and their own approaches to hierarchical Bayesian methods for estimating the parameters of RT models. Very recently, and for the first time, all of the important equations for both the diffusion model and the LBA model have been brought together in a single computer package with coherent programming structure across the models `https://cran.r-project.org/web/packages/rtdists/`. This is a free and open source package for the free and open source statistical language *R* (R Core Team, 2015), and includes joint density and cumulative density function for both models, as well as random sampling functions.

Although the methods for estimating parameters have become increasingly sophisticated, most variants of RT models are relatively complex. Almost all RT models suffer from an identifiability problem (above and beyond the simple scaling problem, see Donkin et al., 2009b). Parameter tradeoffs mean that there are multiple sets of parameter values that can fit data almost equally well. As such, the estimation of the parameters in most RT models requires specifically designed experiments. Typically, multiple within-subject experimental conditions are run, and most RT models require that many of the model's parameters be held constant across those conditions. Even under such conditions, it is important that dozens of trials are collected per condition, though hierarchical approaches can be of particular use when sample sizes are small. With experimental designs less well-suited to RT modeling, parameter estimates should be interpreted with caution.

Theory Development vs. Cognitive Psychometrics. In general, we recommend that researchers err towards using simpler versions of RT models when attempting to do cognitive psychometrics. It is highly likely that certain assumptions in more complex RT models are true. For example, no one would question that there is trial-to-trial variability in the time to make a motor response once a decision is made. Further, as we increase the quality of our data, our models of decision-making are likely to become increasingly complex. Therefore, in terms of theory development, more complex models are inevitable.

It is important to keep in mind, however, the distinction between a model whose purpose is the development of theory, and a model who purpose is measurement. Our conjecture is that the more complex aspects of behavior are not reliably identifiable in typical experiments (i.e., those not specifically designed to measure such processes). When such complexity is not present in the data, then the models will tend to over-fit, and thus yield less reliable parameter estimates. As such, we suggest that models with fewer parameters, and fewer assumptions, are more appropriate tools for cognitive psychometrics. For example, a hierarchical Bayesian implementation of a diffusion model that excludes all forms of between-trial variability (c.f., Wabersich & Vandekerckhove, 2014) can be used in impressively complex applications (Vandekerckhove, 2014), as can the simple linear ballistic accumulator model (Jones, Hawkins, & Brown, 2015).

Model Selection

A related statistical issue concerns how one decides which experimental manipulations influence which model parameters. For example, how does one decide whether it is drift rate, response thresholds, or non-decision processes that differ across the factors of an experiment? There are many approaches to dealing with this issue. One common method is to estimate the drift rate, threshold, and non-decision parameters freely, and use a null-hypothesis statistical testing to determine whether there exist any differences in those parameters across conditions (e.g., Ratcliff, Thapar, Gomez, & McKoon, 2004; Voss, Rothermund, & Voss, 2004). Given the known issues with both null hypothesis testing and parameter estimation for RT models, this approach can be problematic.

Another common approach is to treat the question as a model selection problem. The question is whether model A, which is one particular parameterization of an RT model, gives a more parsimonious account of the data than model B, an alternative parameterization of the same model. The two parameterizations might differ in whether they allow drift rate to differ between the experimental conditions, or threshold to vary, for example. Standard model selection approaches like the Akaike and Bayesian Information Criteria are easy to use, but carry with them their own respective issues, such as being too lenient or punitive with respect to model complexity. It is often useful to carry out bootstrapping simulation studies to determine which of these criteria are appropriate (see Wagenmakers, Ratcliff, Gomez, & Iverson, 2004).

Ideally, one would use more principled model selection techniques such as minimum description length, or Bayes factors (Myung, 2000). At the moment, such approaches are too computationally expensive for RT models. At present, computational shortcuts, such as the Savage-Dickey test (Wagenmakers, Lodewyckx, Kuriyal, & Grasman, 2010), allow us to estimate Bayes factors for nested models. However, in our experience, these shortcuts have not been quite as reliable as hoped. Cross validation methods have been very useful, but come at a substantial cost in terms of computational time. Cross validation for an RT model usually involves leaving out a small fraction of each subject's data, then fitting the model to the remaining data. The fitted model is then compared to the left-out data and a goodness-of-fit calculated. This procedure is repeated several times, with different sets of left-out data, and results averaged. The average goodness-of-fit to the left-out data provides an easy way to compare different models, without relying on precise parameter estimation, and while being sensitive to model complexity. One ongoing practical issue with cross validation concerns the relative sizes of the calibration and validation data sets. This choice creates a bias-variance tradeoff, with no one-size-fits-all solution.

Model Fit

An important assumption of any cognitive psychometric use of an RT model is that the model adequately fits the data. The principle is that one should only ruly upon the inferences from an RT model if it adequately mimics the observed data. Unfortunately, there are relatively few good methods for assessing the quality of agreement between observed data and the predictions of the RT model (i.e., given a particular set of parameters, or distribution of parameters).

Currently, the standard approach is to plot the model predictions alongside the observed data and ask whether the model is doing a "good enough" job. The difficulty, of course, is how one determines what qualifies as good enough. One approach is to find a version of the RT model that has enough parameters that it gives a near perfect account of the data. The idea is that this more complex model is almost certainly over-fitting the data. If the simpler parameterization provides a more parsimonious account of the data than the saturated model, according to one or more model selection metrics, then one can argue that the simpler version of the model fits sufficiently well.

It is worth noting that again the distinction between assessing fit for the purpose of theory development and for the purpose of cognitive psychometrics. From a psychometric perspective, provided that the most reliable and important features of the data are captured, it is probably safe to draw inferences from simpler models, even though they may not capture the full spectrum of observed data patterns (e.g., the relative speed of correct and error responses). From the perspective of theory development, however, it seems much more important that all data patterns are captured, whenever they are demonstrably reliable. Often times, it will simply come down to the quality of the data. Generally speaking, the data collected to develop and test theory is of much higher quality than that collected for typical cognitive psychometric applications. As such, many of the caveats we discuss relating to theory development and cognitive psychometrics follow directly from considerations of model parsimony and avoiding over-fitting.

Summary

RT data, especially those arising from repeated simple decisions, continue to be extremely informative in a very wide variety of psychological research fields. It can be misleading to separately analyze mean RT and accuracy, and so the past fifty years has seen the development of sophisticated decision-making theories that allow joint analysis of the two measures. These theories are based on the idea that evidence about the decision accumulates over time, and a decision is triggered when a sufficient amount of evidence is gathered in favor of one choice over another. Evidence accumulation models have proven extremely successful, both as mechanistic explanations of the cognitive processes underlying decision-making, and as tools for the estimation of cognitive components contributing to observed effects. The models have been applied to data from a very wide array of experiments, in both applied and basic research.

Recent work has also linked the process of evidence accumulation with neural processes which might support decision-making behavior, and with analyses of statistical optimality which might explain the goals decision-making behavior. The links with neural data have been made very detailed by neuroimaging of human decision-makers, and electrophysiological recordings from non-human primate decision-makers. The early theories of neural mechanisms of decision-making bore many similarities to the early cognitive theories of decision-making, and these similarities have been explored in detail since, leading to well-unified cross-disciplinary accounts. Statistical theories of optimality in decision-making are also similar to early cognitive accounts of decision-making, but subsequent investigation of the similarity has not proven quite as fruitful as in neuroscience.

For many years, the routine application of evidence accumulation models to data was made difficult by the mathematical and computational problems involved in parameter estimation. More recently, these barriers to use have been reduced, by the development of simpler models and of more user-friendly and general-purpose analysis software. These developments have created a large and diverse community of researchers who analyze RTs using evidence accumulation models, and who further develop the models themselves, from very different perspectives. With such support, we anticipate a bright future for decision-making research.

BOXES

BOX: How to plot choice RT data

The data from a single condition in a decision-making experiment form a joint distribution over response choice and RT. That is, there are separate RT distributions for each response choice, but these distributions are of different sizes, such that their area adds up to one, across all different responses. Figure 3 provides three common ways to visualize the data from a single condition within a typical experiment. To create the figures, we simulated data to mimic performance in a standard two-choice experiment. This data may represent the behavior of a single individual who made one response on approximately 80% of trials, and took about 750 ms to respond on average.

The leftmost plot shows this simulated data as a pair of histograms. To create this histogram, the RT data for each response were binned into 50 ms chunks. The dominant response is plotted in green, and the less frequent response in red. The main advantage of histograms is that they are easy to interpret. We can immediately see the positive skew of the RT distribution, and the relative frequency of the two responses is fairly clear – there are many more responses in the green distribution than the red. However, histograms are rarely used to compare the predictions of RT models with observed data. The three main disadvantages of histograms are: (a) it is easy to hide discrepancies between a model and data, due to the flexibility permitted when choosing the size of the bins; (b) they can make very complex plots, if there are many different experimental conditions to display; and (c) it is difficult to present aggregated data. For example, if one were to plot the distribution of all individuals' RTs as a histogram, there is no guarantee that the shape of the histogram would reflect the properties of the individuals.

The center plot is a cumulative distribution function plot (CDF). These plots provide an efficient means of simultaneously illustrating accuracy and the shape of the correct and

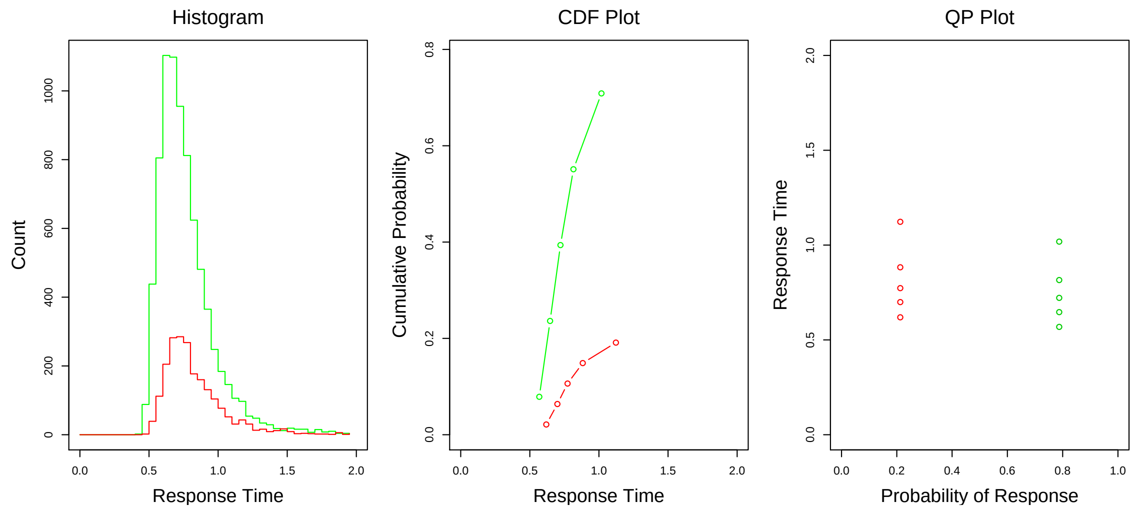

Figure 3. Simulated data from a two-choice experiment are plotted in three different, but common, methods. The details of these plots, and their relative merits and drawbacks are discussed in text.

incorrect RT distributions. Each plot is made up of quantile estimates from the two RT distributions. The quantile estimates show the RT below which 10%, 30%, 50%, 70% and 90% of the responses in that distribution fall. The positions of the quantiles on the x-axis reflect the speed at which responses are made, so that slower distributions stretch further to the right. The heights of the functions indicate, separately for each response, the absolute cumulative proportion of responses with RTs below the quantile cutoff. So, as a particular response becomes more dominant, the distance between the green and red functions increases. CDF plots are more difficult for some people to read than histograms, but they support averaging across participants very well (the quantiles are calculated for each participant, and those are averaged).

Finally, the rightmost plot is a quantile-probability plot (QP), which plots exactly the same summary statistics as the CDF plot, but in a different way. QP plots are an efficient way of displaying the important information from a set of choice RT data the horizontal axis contains response probability (accuracy) information and the vertical axis contains information about the RT distribution. Unlike the CDF plot, the quantiles of the RT distributions are plotted above one another, and the accuracy information is given by the position of the quantiles on the horizontal axis. One advantage of QP plots over CDF plots is that results for more than one condition can be given in the same graph. This often works well when the conditions differ sufficiently in accuracy.

Both CDF and QP plots easily permit comparison of group-level model predictions and data. Group QP or cumulative probability plots can be obtained by averaging quantiles and probabilities for each individual, also have the advantage that they tend to be more representative of individual results (e.g., such averages do not suffer from the problems that occur with histograms Rouder & Speckman, 2004). To represent the model predictions on these plots at the group level, one calculates the model's predicted quantiles for each individual and averages these together in the same way as the data. This means that we apply the same averaging process to create summary information for model predictions as for the data, and so both summaries are subjected equally to any distorting effects of averaging.

BOX: Some application areas

Evidence accumulation models of choice RT are increasingly used to examine the psychological processes underlying rapid decisions. Since the parameters of evidence accumulation models quantify different aspects of the decision process, variations among experimental conditions in model parameters can provide insights into latent psychological processes beyond those available from traditional measures. Theories based on the idea of evidence accumulation have been successfully applied to many different paradigms, including: simple perceptual decisions (Usher & McClelland, 2001), visual short-term memory (P. L. Smith & Ratcliff, 2009), absolute identification (Brown et al., 2008), lexical decision (Ratcliff, Gomez, & McKoon, 2004; Wagenmakers et al., 2008), and the neural correlates of behavioral measures (Farrell, Ratcliff, Cherian, & Segraves, 2006; B. U. Forstmann et al., 2008; Ho et al., 2009).

Evidence accumulation models have been used as tools for the measurement of cognitive processing (see the section on "cognitive psychometrics") in a vast array of different paradigms, including: consumer choice (Busemeyer & Townsend, 1992; Hawkins et al., 2014); understanding the cognition of people with depression (White, Ratcliff, Vasey, & McKoon, 2009; Ho et al., 2014); personality traits (Vandekerckhove, 2014); pain sensitivity (Wiech et al., 2014); car driving (Ratcliff, 2015); video game pratice effects (van Ravenzwaaij, Boekel, Forstmann, Ratcliff, & Wagenmakers, 2014); psychopharmacology (Winkel et al., 2012); and many others.

Evidence accumulation models have traditionally been developed for, and applied to, very simple decision tasks – decisions that take less than a second to make, about single-attribute stimuli such as luminosity, loudness, motion, or orientation. In recent years, evidence accumulation models have been extended to much more sophisticated decision-making scenarios, including:

- Multi-attribute choices, such as are frequently faced by consumers, where products vary on price, quality, availability, looks, and many other attributes (Busemeyer & Townsend, 1992; Trueblood, Brown, & Heathcote, 2014; Krajbich & Rangel, 2011).
- Decisions with more complicated response mappings. The standard decision task has a simple one-to-one mapping between stimuli and reponses ("press the left button if the stimulus is blue"), but many interesting tasks have more complex response rules, such as the go/no-go task, the stop-signal task, and the redundant signals task. Evidence accumulation

models have recently been extended to all of these (Gomez, Ratcliff, & Perea, 2007; Matzke, Love, & Heathcote, 2015; Palada et al., n.d.; Eidels, Donkin, Brown, & Heathcote, 2010; Donkin, Little, & Houpt, 2014; Houpt, Townsend, & Donkin, 2014; Endres & Finn, 2014).

- Decisions involving more than one response for each choice, such as "best-worst scaling" tasks (Hawkins et al., in press)
- Tasks in which responses may come from a mixture of latent processes, such as slot-based models of visual working memory (Donkin, Nosofsky, Gold, & Shiffrin, 2013; Nosofsky & Donkin, 2016), or from more complex rules (Fific, Little, & Nosofsky, 2010; Little, Nosofsky, & Denton, 2011; Little, Nosofsky, Donkin, & Denton, 2013).

BOX: How the diffusion model works.

In the diffusion model (Ratcliff, 1978; Ratcliff & Rouder, 2000; Wagenmakers, 2009; van Ravenzwaaij & Oberauer, 2009), stimulus processing is conceptualized as the noisy accumulation of evidence over time. A response is initiated when the accumulated evidence reaches a predefined threshold (Figure 4).

The diffusion model applies to tasks in which the participant has to decide quickly between two alternatives. For instance, in a *lexical decision task*, participants have to decide whether a letter string is a valid word, such as RUN, or a nonword, such as NUR. The RTs in such tasks generally do not exceed 1.0 or 1.5 seconds. The four key parameters of the diffusion model are (1) the speed of information processing, quantified by mean drift rate v; (2) response caution, quantified by boundary separation a; (3) a priori bias, quantified by mean starting point z; and (4) mean non–decision time, quantified by T_{er}.

The model assumes that the decision process starts at z, after which information is accumulated with a signal–to–noise ratio that is governed by mean drift rate v.[1] Conceptually, drift rate captures a range of factors that affect information accumulation, including characteristics of the stimuli, the task, and the participant. Small drift rates (near $v = 0$) produce long RTs and high error rates. Boundary separation (a) determines the speed–accuracy tradeoff; lowering boundary separation leads to faster RTs at the cost of a higher error rate. A starting point of $z = .5a$ indicates an unbiased decision process. Together, these parameters generate a distribution of decision times DT. The observed RT, however, also consists of stimulus–nonspecific components such as response preparation and motor execution, which together make up non–decision time T_{er}. The model assumes that non–decision time T_{er} simply shifts the distribution of DT, such that $RT = DT + T_{er}$ (Luce, 1986). The full diffusion model includes parameters that specify across–trial variability in drift rate, starting point, and non–decision time (Ratcliff & Tuerlinckx, 2002).

[1]Mathematically, the change in evidence X is described by a stochastic differential equation $dX(t) = v \cdot dt + s \cdot dW(t)$, where $W(t)$ represents the Wiener noise process (i.e., idealized Brownian motion). Parameter s represents the standard deviation of $dW(t)$ and is usually fixed.

BOX: How the LBA model works.

Figure 5 illustrates decision processing in a pair of LBA units. Suppose that the figure represents a decision about whether a cloud of dots appears to be moving to the left or to the right, requiring a "left" or "right" response, respectively. Presentation of the stimulus causes evidence to accumulate for both the "left" and "right" responses separately, as indicated by the two lines (one solid and one dotted) in Figure 5. The vertical axis of the figure represents the amount of evidence that has been accumulated, and the horizontal axis shows how much decision time has passed. The amount of evidence in each accumulator increases linearly until one reaches the response threshold, and the decision time is the time taken for the first accumulator to reach threshold. The predicted RT is made up of the decision time plus a non-decision time, quantified by parameter $T_e r$.

The slopes of the lines in Figure 5 indicate the rates at which evidence is accumulated for each response, and are usually referred to as the drift rates. If the physical stimulus favors a "left" response, the drift rate for the "left" response accumulator will usually be larger than for the "right" response accumulator. Drift rates are assumed to be set by physical stimulus properties and by the demands of the task. For example, in the random dot motion task, decisions might be made easier by making the displayed dots drift more steadily in one direction. This would provide stronger evidence that "left" was the correct response, and the drift rate for that response would increase. Drift rates are also assumed to be modulated by sensory and attentional processing, and the overall efficiency of the cognitive system. For example, Schmiedek, Oberauer, Wilhelm, Süß, and Wittmann (2007) found larger drift rates for participants with higher working memory capacity and fluid intelligence. In the LBA, there are two different drift rates: one for each accumulator (corresponding to "left" and "right" responses). The relative size of drift rate parameters describes differences in task performance between different conditions or groups. Although not explicitly illustrated in Figure 5, drift rates in the LBA are assumed to vary randomly from trial-to-trial according to a normal distribution with mean v and standard deviation s, reflecting trial-to-trial fluctuations in factors such as attention and arousal.

The amount of evidence in each accumulator before the beginning of the decision process also varies from trial-to-trial. The starting evidence for each accumulator is assumed to follow a uniform distribution whose minimum value is set (without loss of generality) at zero evidence for all accumulators, and whose upper value is determined by a parameter A. Hence, the average amount (across trials) of evidence in each accumulator before accumulation begins is $\frac{A}{2}$. The height of the response threshold that must be reached is called b, and is represented by the horizontal dotted line in Figure 5. The value of b relative to the average starting activation ($\frac{A}{2}$), provides a measure of average response caution, because the difference ($b - \frac{A}{2}$) is the average amount of evidence that must accumulate before a response will be triggered. In Figure 5, the same response threshold (b) is used for both accumulators; this indicates that the same amount of evidence is required, on average, before either response is made. If participants choose to favor one particular response (i.e., a response bias), b and/or A might be smaller for the preferred response. Response bias leads to a speed-accuracy trade-off, as the preferred response is made more quickly, but it is also made more often when incorrect, reducing accuracy.

The time taken for each accumulator in the LBA to reach threshold on any given trial is the distance between the response threshold and the start point of activation, divided by the rate of evidence accumulation. The observed decision time on any given trial, however, is the time for the fastest accumulator to reach threshold. The formula for the distribution across trials of the time taken for the fastest accumulator to reach threshold is given by Brown and Heathcote (2008); Terry et al. (2015). This formula makes it possible to estimate the model's parameters from data.

The original formulation of the LBA model, described above, assumed normal distributions for the variability in drift rates from trial to trial. This creates a conceptual problem because it necessarily means that some drift rates, on some trials, will be negative, potentially leading to undefined RTs. Although this problem has not so far proven practically important, it has been addressed in recent work by Terry et al. (2015). This work has shown how the analytic tractability of the LBA model can be maintained even when using a variety of different drift rate distributions which are all constrained to positive values only (such as the gamma and lognormal distributions).

END BOXES

References

Audley, R., & Pike, A. (1965). Some alternative stochastic models of choice1. *British Journal of Mathematical and Statistical Psychology*, *18*(2), 207–225.

Batchelder, W. H. (1998). Multinomial processing tree models and psychological assessment. *Psychological Assessment*, *10*, 331–344.

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review*, *6*, 57–86.

Belliveau, J., Kennedy, D., McKinstry, R., Buchbinder, B., Weisskoff, R., Cohen, M., … Rosen, B. (1991). Functional mapping of the human visual cortex by magnetic resonance imaging. *Science*, *254*, 716–719.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two–alternative forced choice tasks. *Psychological Review*, *113*, 700–765.

Bowman, N. E., Kording, K. P., & Gottfried, J. A. (2012). Temporal integration of olfactory perceptual evidence in human orbitofrontal cortex. *Neuron*, *75*, 916–927.

Brainard, R. W., Irby, T. S., Fitts, P. M., & Alluisi, E. A. (1962). Some variables influencing the rate of gain of information. *Journal of Experimental Psychology*, *63*, 105–110.

Brenner, D., Williamson, S., & Kaufman, L. (1975). Visually evoked magnetic fields of the human brain. *Science*, *190*, 480–482.

Brown, S. D., & Heathcote, A. (2005). A ballistic model of choice response time. *Psychological Review*, *112*, 117–128.

Brown, S. D., & Heathcote, A. J. (2008). The simplest complete model of choice reaction time: Linear ballistic accumulation. *Cognitive Psychology*, *57*, 153–178.

Brown, S. D., Marley, A., Donkin, C., & Heathcote, A. J. (2008). An integrated model of choices and response times in absolute identification. *Psychological Review*, *115*(2), 396–425.

Brown, S. D., Steyvers, M., & Wagenmakers, E.-J. (2009). Observing evidence accumulation during multi–alternative decisions. *Journal of Mathematical Psychology*, *53*, 453–462.

Busemeyer, J. R., & Townsend, J. T. (1992). Fundamental derivations from decision field theory. *Mathematical Social Sciences*, *23*, 255–282.

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic–cognitive approach to decision making. *Psychological Review*, *100*, 432–459.

Busemeyer, J. R., Wang, Z., & Townsend, J. T. (2006). Quantum dynamics of human decision-making. *Journal of Mathematical Psychology*, *50*(3), 220–241.

Carpenter, R. H. S., & Reddi, B. A. J. (2001). Deciding between the deciders: Two models of reaction time may happily coexist. *Nature Neuroscience*, *4*, 337.

Cavanagh, J. F., Wiecki, T. V., Cohen, M. X., Figueroa, C. M., Samanta, J., Sherman, S. J., & Frank, M. J. (2011, September). Subthalamic nucleus stimulation reverses mediofrontal influence over decision threshold. *Nature Publishing Group*, *14*(11), 1462–1467.

Churchland, A. K., Kiani, R., & Shadlen, M. N. (2008). Decision-making with multiple alternatives. *Nature Neuroscience*, *11*, 693–702.

Cisek, P., Puskas, G. A., & El-Murr, S. (2009). Decisions in changing conditions: The urgency–gating model. *The Journal of Neuroscience*, *29*(37), 11560–11571.

Dassonville, P., Lewis, S. M., Foster, H., & Ashe, J. (1999). Choice and stimulus–response compatibility affect duration of response selection. *Cognitive Brain Research*, *7*, 235–240.

de Hollander, G., Forstmann, B. U., & Brown, S. D. (2015). Different ways of linking behavioral and neural data via computational cognitive models. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*.

Ditterich, J. (2006a). Evidence for time–variant decision making. *European Journal of Neuroscience*, *24*, 3628–3641.

Ditterich, J. (2006b). Stochastic models of decisions about motion direction: Behavior and physiology. *Neural Networks*, *19*, 981–1012.

Donkin, C., Brown, S. D., & Heathcote, A. (2009a). Getting more from accuracy and response time data: Methods for fitting the linear ballistic accumlator model. *Behavior Research Methods*, *41*, 1095–1110.

Donkin, C., Brown, S. D., & Heathcote, A. (2009b). The overconstraint of response time models: Rethinking the scaling problem. *Psychonomic Bulletin and Review*, *16*(6), 1129–1135.

Donkin, C., Little, D., & Houpt, J. (2014). Assessing the speed-accuracy trade-off effect on the capacity of information processing. *Journal of Experimental Psychology: Human Perception and Performance*, *40*, 1183-1202.

Donkin, C., Nosofsky, R. M., Gold, J., & Shiffrin, R. M. (2013). Discrete-slots models of visual working memory response times. *Psychological Review*, *120*, 873–902.

Drugowitsch, J., Moreno-Bote, R., Churchland, A. K., Shadlen, M. N., & Pouget, A. (2012). The cost of accumulating evidence in perceptual decision making. *Journal of Neuroscience*, *32*, 3612–3628.

Eidels, A., Donkin, C., Brown, S. D., & Heathcote, A. (2010). Converging measures of workload capacity. *Psychonomic Bulletin & Review*.

Endres, C., M. Donkin, & Finn, P. (2014). An information processing/associative learning account of behavioral disinhibition in externalizing psychopathology. *Experimental Clinical Psychopharmacology*, *22*, 122–132.

Farrell, S., Ratcliff, R., Cherian, A., & Segraves, M. (2006). Modeling unidimensional categorization in monkeys. *Learning and Behavior*, *34*, 86–101.

Feller, W. (1971). *An introduction to probability theory and its applications: Vol. ii.* New York: John Wiley & Sons.

Fific, M., Little, D., & Nosofsky, R. M. (2010). Logical-rule models of classification response times: A synthesis of mental-architecture, random-walk, and decision-bound approaches. *Psychological Review*, *117*, 309–348.

Forstmann, B., Ratcliff, R., & Wagenmakers, E.-J. (2016). Sequential sampling models in cognitive neuroscience: Advantages, applications, and extensions. *Annual review of psychology*, *67*, 641–666.

Forstmann, B. U., Anwander, A., Schäfer, A., Neumann, J., Brown, S. D., Wagenmakers, E.-J., ... Turner, R. (2010). Cortico-striatal connections predict control over speed and accuracy in perceptual decision making. *Proceedings of the National Academy of Sciences*, *107*(36), 15916–15920.

Forstmann, B. U., Dutilh, G., Brown, S., Neumann, J., von Cramon, D. Y., Ridderinkhof, K. R., & Wagenmakers, E.-J. (2008). Striatum and pre–SMA facilitate decision–making under time pressure. *Proceedings of the National Academy of Sciences*, *105*, 17538–17542.

Frank, M. J. (2006). Hold your horses: A dynamic computational role for the subthalamic nucleus in decision making. *Neural Networks*, *19*.

Frank, M. J., & Claus, E. D. (2006). Anatomy of a decision: Striato–orbitofrontal interactions in reinforcement learning, decision making, and reversal. *Psychological Review*, *113*, 300–326.

Gold, J. I., & Shadlen, M. N. (2001, January). Neural computations that underlie decisions about sensory stimuli. *Trends in Cognitive Sciences*, *5*(1), 10–16.

Gomez, P., Ratcliff, R., & Perea, M. (2007). A model of the go/no-go task. *Journal of Experimental Psychology: General*, *136*(3), 389.

Grasman, R. P., Wagenmakers, E.-J., & van der Maas, H. L. (2009). On the mean and variance of response times under the diffusion model with an application to parameter estimation. *Journal of Mathematical Psychology*, *53*(2), 55–68.

Gronlund, S. D., & Ratcliff, R. (1991). Analysis of the Hockley and Murdock decision model. *Journal of Mathematical Psychology*, *35*, 319–344.

Hanes, D. P., & Schall, J. D. (1996, October). Neural control of voluntary movement initiation. *Science*, *274*(5286), 427–430.

Hawkins, G. E., Brown, S. D., Steyvers, M., & Wagenmakers, E.-J. (2012). An optimal adjustment procedure to minimize experiment time in decisions with multiple alternatives. *Psychonomic Bulletin & Review*, *19*, 339–348.

Hawkins, G. E., Forstmann, B. U., Wagenmakers, E.-J., Ratcliff, R., & Brown, S. D. (2015). Revisiting the evidence for collapsing boundaries and urgency signals in perceptual decision-making. *The Journal of Neuroscience*, *35*(6), 2476–2484.

Hawkins, G. E., Marley, A. A. J., Heathcote, A., Flynn, T. N., Louviere, J. J., & Brown, S. D. (2014). Integrating cognitive process and descriptive models of attitudes and preferences. *Cognitive Science*, *38*, 701–735.

Hawkins, G. E., Marley, A. A. J., Heathcote, A., Flynn, T. N., Louviere, J. J., & Brown, S. D. (in press). The best of times and the worst of times are interchangeable. *Decision*.

Hawkins, G. E., Wagenmakers, E.-J., Ratcliff, R., & Brown, S. D. (2015). Discriminating evidence accumulation from urgency signals in speeded decision making. *Journal of neurophysiology*, jn–00088.

Hick, W. E. (1952). On the rate of gain of information. *Quarterly Journal of Experimental Psychology*, *4*, 11–26.

Hillyard, S. A., Hink, R. F., Schwent, V. L., & Picton, T. W. (1973). Electrical signs of selective attention in the human brain. *Science*, *182*, 177–180.

Ho, T. C., Brown, S., & Serences, J. T. (2009). Domain general mechanisms of perceptual decision making in human cortex. *The Journal of Neuroscience*, *29*(27), 8675–8687.

Ho, T. C., Yang, G., Wu, J., Cassey, P., Brown, S. D., Hoang, N., ... others (2014). Functional connectivity of negative emotional processing in adolescent depression. *Journal of affective disorders*, *155*, 65–74.

Hockley, W. E., & Murdock, B. B. (1987). A decision model for accuracy and response latency in recognition memory. *Psychological Review*, *94*, 341–358.

Houpt, J., Townsend, J. T., & Donkin, C. (2014). A new perspective on visual word processing efficiency. *Acta Psychologica*, *145*, 118-127.

Hyman, R. (1953). Stimulus information as a determinant of reaction time. *Journal of Experimental Psychology*, *45*, 188–196.

Jones, L. G., Hawkins, G. E., & Brown, S. D. (2015). Using best–worst scaling to improve psychological service delivery: An innovative tool for psychologists in organized care settings. *Psychological services*, *12*(1), 20.

Krajbich, I., & Rangel, A. (2011). A multi–alternative drift diffusion model predicts the relationship between visual fixations and choice in value–based decisions. *Proceedings of the National Academy of Science*, *108*, 13852–13857.

Kveraga, K., Boucher, L., & Hughes, H. C. (2002). Saccades operate in violation of Hick's Law. *Experimental Brain Research*, *146*, 307–314.

LaBerge, D. A. (1962). A recruitment theory of simple behavior. *Psychometrika*, *27*(4), 375–396.

LaBerge, D. A. (1994). Quantitative models of attention and response processes in shape identification tasks. *Journal of Mathematical Psychology*, *38*, 198–243.

Lacouture, Y., & Marley, A. A. J. (1995). A mapping model of bow effects in absolute identification. *Journal of Mathematical Psychology*, *39*, 383–395.

Laming, D. R. J. (1968). *Information theory of choice–reaction times.* London: Academic Press.

Laursen, A. M. (1977). Task dependence of slowing after pyramidal lesions in monkeys. *Journal of Comparative and Physiological Psychology*, *91*, 897–906.

Lee, K.-M., Keller, E. L., & Heinen, S. J. (2005). Properties of saccades generated as a choice response. *Experimental Brain Research*, *162*, 278–286.

Lee, M. D., & Wagenmakers, E. J. (2014). *Bayesian cognitive modeling.* Cambridge University Press.

Leite, F. P., & Ratcliff, R. (2010). Modeling reaction time and accuracy of multiple–alternative decisions. *Attention, Perception & Psychophysics*, *72*, 246–273.

Link, S. W., & Heath, R. A. (1975). A sequential theory of psychological discrimination. *Psychometrika*, *40*, 77–105.

Little, D., Nosofsky, R. M., & Denton, S. (2011). Response-time tests of logical-rule models of categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Little, D., Nosofsky, R. M., Donkin, C., & Denton, S. (2013). Logical rules and the classification of integral-dimension stimuli. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *39*, 801–820.

Lo, C.-C., & Wang, X.-J. (2006). Cortico-basal ganglia circuit mechanism for a decision threshold in reaction time tasks. *Nature Neuroscience*, *9*, 956–963.

Luce, R. D. (1986). *Response times*. New York: Oxford University Press.

Matzke, D., Love, J., & Heathcote, A. (2015). A bayesian approach for estimating the probability of trigger failures in the stop-signal paradigm. *Behavior Research Methods*.

Matzke, D., & Wagenmakers, E.-J. (2009). Psychological interpretation of ex–Gaussian and shifted Wald parameters: A diffusion model analysis. *Psychonomic Bulletin & Review*, *16*, 798–817.

Merkle, E., Smithson, M., & Verkuilen, J. (2011). Hierarchical models of simple mechanisms underlying confidence in decision making. *Journal of Mathematical Psychology*, *55*, 57–67.

Milosavljevic, M., Malmaud, J., Huth, A., Koch, C., & Rangel, A. (2010). The drift diffusion model can account for the accuracy and reactime of value–based choices under high and low time pressure. *Judgment and Decision Making*, *5*, 437–449.

Moran, R., Teodorescu, A. R., & Usher, M. (2015). Post choice information integration as a causal determinant of confidence: Novel data and a computational account. *Cognitive psychology*, *78*, 99–147.

Mulder, M. J., Wagenmakers, E.-J., Ratcliff, R., Boekel, W., & Forstmann, B. U. (2012, February). Bias in the Brain: A Diffusion Model Analysis of Prior Probability and Potential Payoff. *Journal of Neuroscience*, *32*(7), 2335–2343.

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, *44*, 190–204.

Nosofsky, R. M., & Donkin, C. (2016). Response-time evidence for mixed memory states in a sequential-presentation change-detection task. *Cognitive Psychology*, *84*, 31–62.

Pachella, R. G. (1974). The interpretation of reaction time in information–processing research. In B. H. Kantowitz (Ed.), *Human information processing: Tutorials in performance and cognition* (pp. 41–82). Hillsdale (NJ): Lawrence Erlbaum Associates.

Pachella, R. G., & Fisher, D. (1972). Hick's Law and the speed–accuracy trade–off in absolute judgment. *Journal of Experimental Psychology*, *92*, 378–384.

Palada, H., Neal, A., Vuckovic, A., Martin, R., Samules, K., & Heathcote, A. (n.d.). Evidence accumulation in a complex task: Making choices about concurrent multiattribute stimuli under time pressure. *Journal of Experimental Psychology: Applied*.

Pietsch, A., & Vickers, D. (1997). Memory capacity and intelligence: Novel techniques for evaluating rival models of a fundamental information processing mechanism. *The Journal of General Psychology*, *124*, 231–339.

Pike, A. R. (1966). Stochastic models of choice behaviour: Response probabilities and latencies of finite Markov chain systems. *British Journal of Mathematical and Statistical Psychology*, *21*, 161–182.

Pike, A. R. (1973). Response latency models for signal detection. *Psychological Review*, *80*, 53–68.

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. *Psychological review*, *117*(3), 864.

Purcell, B. A., Heitz, R. P., Cohen, J. Y., Schall, J. D., Logan, G. D., & Palmeri, T. J. (2010). Neurally constrained modeling of perceptual decision making. *Psychological Review*, *117*, 1113–1143.

Purcell, B. A., Schall, J. D., Logan, G. D., & Palmeri, T. J. (2012, Mar). From salience to saccades: multiple-alternative gated stochastic accumulator model of visual search. *Journal of Neuroscience*, *32*(10), 3433–3446. Retrieved from `http://dx.doi.org/10.1523/JNEUROSCI.4622-11.2012` doi: 10.1523/JNEUROSCI.4622-11.2012

R Core Team. (2015). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59–108.

Ratcliff, R. (2015). Modeling one-choice and two-choice driving tasks. *Attention, Perception, & Psychophysics*, 1–11.

Ratcliff, R., Gomez, P., & McKoon, G. (2004). Diffusion model account of lexical decision. *Psychological Review*, *111*, 159–182.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two–choice decision tasks. *Neural Computation*, *20*, 873–922.

Ratcliff, R., Philiastides, M. G., & Sajda, P. (2009, April). Quality of evidence for perceptual decision making is indexed by trial-to-trial variability of the EEG. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(16), 6539–6544.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two–choice decisions. *Psychological Science*, *9*, 347–356.

Ratcliff, R., & Rouder, J. N. (2000). A diffusion model account of masking in two–choice letter identification. *Journal of Experimental Psychology: Human Perception and Performance*, *26*, 127–140.

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two–choice reaction time. *Psychological Review*, *111*, 333–367.

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in cognitive sciences*, *20*(4), 260–281.

Ratcliff, R., & Starns, J. J. (2009). Modeling confidence and response time in recognition memory. *Psychological review*, *116*(1), 59.

Ratcliff, R., & Starns, J. J. (2013). Modeling confidence judgments, response times, and multiple choices in decision making: recognition memory and motion discrimination. *Psychological review*, *120*(3), 697.

Ratcliff, R., Thapar, A., Gomez, P., & McKoon, G. (2004). A diffusion model analysis of the effects of aging in the lexical–decision task. *Psychology and Aging*, *19*, 278–289.

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review*, *9*, 438–481.

Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review*, *102*, 261–300.

Rouder, J. N., & Speckman, P. L. (2004). An evaluation of the vincentizing method of forming group-level response time distributions. *Psychonomic Bulletin & Review*, *11*(3), 419–427.

Schall, J. D. (2001). Neural basis of deciding, choosing, and acting. *Nature Reviews Neuroscience*, *2*, 33–42.

Schmiedek, F., Oberauer, K., Wilhelm, O., Süß, H.-M., & Wittmann, W. W. (2007). Individual differences in components of reaction time distributions and their relations to working memory and intelligence. *Journal of Experimental Psychology: General*, *136*, 414–429.

Schouten, J. F., & Bekker, J. A. M. (1967). Reaction time and accuracy. *Acta Psychologica*, *27*, 143–153.

Shadlen, M. N., & Newsome, W. T. (1996). Motion perception: Seeing and deciding. *Proceedings of the National Academy of Sciences*, *93*, 628–633.

Smith, J. B., & Batchelder, W. H. (in press). Beta-MPT: Multinomial processing tree models for addressing individual differences. *Journal of Mathematical Psychology*.

Smith, P. L. (2000). Stochastic dynamic models of response time and accuracy: A foundational primer. *Journal of Mathematical Psychology*, *44*, 408–463.

Smith, P. L. (2010, April). Journal of Mathematical Psychology. *Journal of Mathematical Psychology*, *54*(2), 266–283.

Smith, P. L., & Ratcliff, R. (2009). An integrated theory of attention and decision making in visual signal detection. *Psychological Review*, *116*, 283–317.

Smith, P. L., & Vickers, D. (1988). The accumulator model of two–choice discrimination. *Journal of Mathematical Psychology*, *32*, 135–168.

Stone, M. (1960). Models for choice–reaction time. *Psychometrika*, *25*, 251–260.

Sutton, S., Braren, M., Zubin, J., & John, E. (1965). Evoked–potential correlates of stimulus uncertainty. *Science*, *150*, 1187–1188.

Teichner, W. H., & Krebs, M. J. (1974). Laws of visual choice reaction time. *Psychological Review*, *81*, 75–98.

Teller, D. (1984). Linking propositions. *Vision Research*, *24*(10), 1233–1246.

ten Hoopen, G., Akerboom, S., & Raaymakers, E. (1982). Vibrotactual choice reaction time, tactile receptor systems and ideomotor compatibility. *Acta Psychologica*, *50*, 143–157.

Teodorescu, A. R., & Usher, M. (2013). Disentangling decision models: From independence to competition. *Psychological review*, *120*(1), 1.

Terry, A., Marley, A., Barnwal, A., Wagenmakers, E.-J., Heathcote, A., & Brown, S. D. (2015). Generalising the drift rate distribution for linear ballistic accumulators. *Journal of Mathematical Psychology*, *68*, 49–58.

Thura, D., Beauregard–Racine, J., Fradet, C. W., & Cisek, P. (2012). Decision making by urgency gating: Theory and experimental support. *Journal of Neurophysiology*, *108*, 2912–2930.

Townsend, J. T., & Ashby, F. G. (1983). *Stochastic modeling of elementary psychological processes.* London: Cambridge University Press.

Trueblood, J. S., Brown, S. D., & Heathcote, A. (2014). The multiattribute linear ballistic accumulator model of context effects in multialternative choice. *Psychological Review*, *121*, 179–205.

Turner, B. M., Forstmann, B. U., Wagenmakers, E.-J., Brown, S. D., Sederberg, P. B., & Steyvers, M. (2013). A Bayesian framework for simultaneously modeling neural and behavioral data. *Neuroimage*, *72*, 193–206.

Turner, B. M., Sederberg, P. B., Brown, S. D., & Steyvers, M. (2013). A method for efficiently sampling from distributions with correlated dimensions. *Psychological Methods*, *18*, 368–384. doi: 10.1037/a0032222

Turner, B. M., van Maanen, L., & Forstmann, B. U. (2014, May). Informing Cognitive Abstractions through Neuroimaging: The Neural Drift Diffusion Model. *Psychological Review*, 1–66.

Usher, M., & McClelland, J. L. (2001). On the time course of perceptual choice: The leaky competing accumulator model. *Psychological Review*, *108*, 550–592.

Usher, M., Olami, Z., & McClelland, J. (2002). Hick's Law in a stochastic race model with speed–accuracy tradeoff. *Journal of Mathematical Psychology*, *46*, 704–715.

van Ravenzwaaij, D., & Oberauer, K. (2009). How to use the diffusion model: Parameter recovery of three methods: EZ, fast-dm, and DMAT. *Journal of Mathematical Psychology*, *53*, 463–473.

Vandekerckhove, J. (2014). A cognitive latent variable model for the simultaneous analysis of behavioral and personality data. *Journal of Mathematical Psychology*, *60*, 58–71.

Vandekerckhove, J., & Tuerlinckx, F. (2008). Diffusion model analysis with MATLAB: A DMAT primer. *Behavior Research Methods*, *40*.

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical diffusion models for two–choice response times. *Psychological Methods*, *16*, 44–62.

van Ravenzwaaij, D., Boekel, W., Forstmann, B. U., Ratcliff, R., & Wagenmakers, E.-J. (2014). Action video games do not improve the speed of information processing in simple perceptual tasks. *Journal of Experimental Psychology: General*, *143*(5), 1794.

Van Zandt, T. (2000). How to fit a response time distribution. *Psychonomic bulletin & review*, *7*(3), 424–465.

Van Zandt, T., Colonius, H., & Proctor, R. W. (2000). A comparison of two response time models applied to perceptual matching. *Psychonomic Bulletin & Review*, *7*, 208–256.

Verdonck, S., & Tuerlinckx, F. (2014). The ising decision maker: A binary stochastic network for choice response time. *Psychological review*, *121*(3), 422.

Verdonck, S., & Tuerlinckx, F. (2016). Factoring out nondecision time in choice reaction time data: Theory and implications. *Psychological Review*, *123*, 208–218.

Vickers, D. (1979). *Decision processes in visual perception*. London: Academic Press.

Vickers, D., & Lee, M. D. (1998). Dynamic models of simple judgments: I. Properties of a self–regulating accumulator module. *Nonlinear Dynamics, Psychology, and Life Sciences*, *2*, 169–194.

Vickers, D., & Lee, M. D. (2000a). Dynamic models of simple judgments: II. Properties of a self–organizing PAGAN (parallel, adaptive, generalized accumulator network) model for multi–choice tasks. *Nonlinear Dynamics, Psychology, and Life Sciences*, *4*, 1–31.

Vickers, D., & Lee, M. D. (2000b). Dynamic models of simple judgments: II. Properties of a self–organizing PAGAN (parallel, adaptive, generalized accumulator network) model for multi–choice tasks. *Nonlinear Dynamics, Psychology, and Life Sciences*, *4*, 1–31.

Vickrey, C., & Neuringer, A. (2000). Pigeon reaction time, Hick's Law, and intelligence. *Psychonomic Bulletin & Review*, *7*, 284–291.

Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: An empirical validation. *Memory & Cognition*, *32*, 1206–1220.

Voss, A., & Voss, J. (2007). Fast–dm: A free program for efficient diffusion model analysis. *Behavior Research Methods*, *39*, 767–775.

Voss, A., & Voss, J. (2008). A fast numerical algorithm for the estimation of diffusion model parameters. *Journal of Mathematical Psychology*, *52*, 1–9.

Wabersich, D., & Vandekerckhove, J. (2014). Extending jags: A tutorial on adding custom distributions to jags (with a diffusion model example). *Behavior Research Methods*, *46*, 15–28.

Wagenmakers, E.-J. (2009). Methodological and empirical developments for the Ratcliff diffusion model of response times and accuracy. *European Journal of Cognitive Psychology*, *21*, 641–671.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. P. P. P. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the savage-dickey method. *Cognitive Psychology*, *60*, 158–189.

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology*, *48*, 28–50.

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & McKoon, G. (2008). A diffusion model account of criterion shifts in the lexical decision task. *Journal of Memory and Language*, *58*, 140–159.

Wagenmakers, E.-J., van der Maas, H. J. L., & Grasman, R. P. P. P. (2007). An EZ–diffusion model for response time and accuracy. *Psychonomic Bulletin & Review*, *14*, 3–22.

Wald, A., & Wolfowitz, J. (1948). Optimal character of the sequential probability ratio test. *Annals of Mathematical Statistics*, *19*, 326–339.

White, C., Ratcliff, R., Vasey, M. W., & McKoon, G. (2009). Dysphoria and memory for emotional material: A diffusion model analysis. *Cognition and Emotion*, *23*, 181–205.

Wickelgren, W. A. (1977). Speed–accuracy tradeoff and information processing dynamics. *Acta Psychologica*, *41*, 67–85.

Wiech, K., Vandekerckhove, J., Zaman, J., Tuerlinckx, F., Vlaeyen, J. W., & Tracey, I. (2014). Influence of prior information on pain involves biased perceptual decision-making. *Current Biology*, *24*(15), R679–R681.

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. *Frontiers in neuroinformatics*, *7*, 14.

Winkel, J., Van Maanen, L., Ratcliff, R., Van der Schaaf, M. E., Van Schouwenburg, M. R., Cools, R., & Forstmann, B. U. (2012). Bromocriptine does not alter speed–accuracy tradeoff. *Frontiers in neuroscience*, *6*.

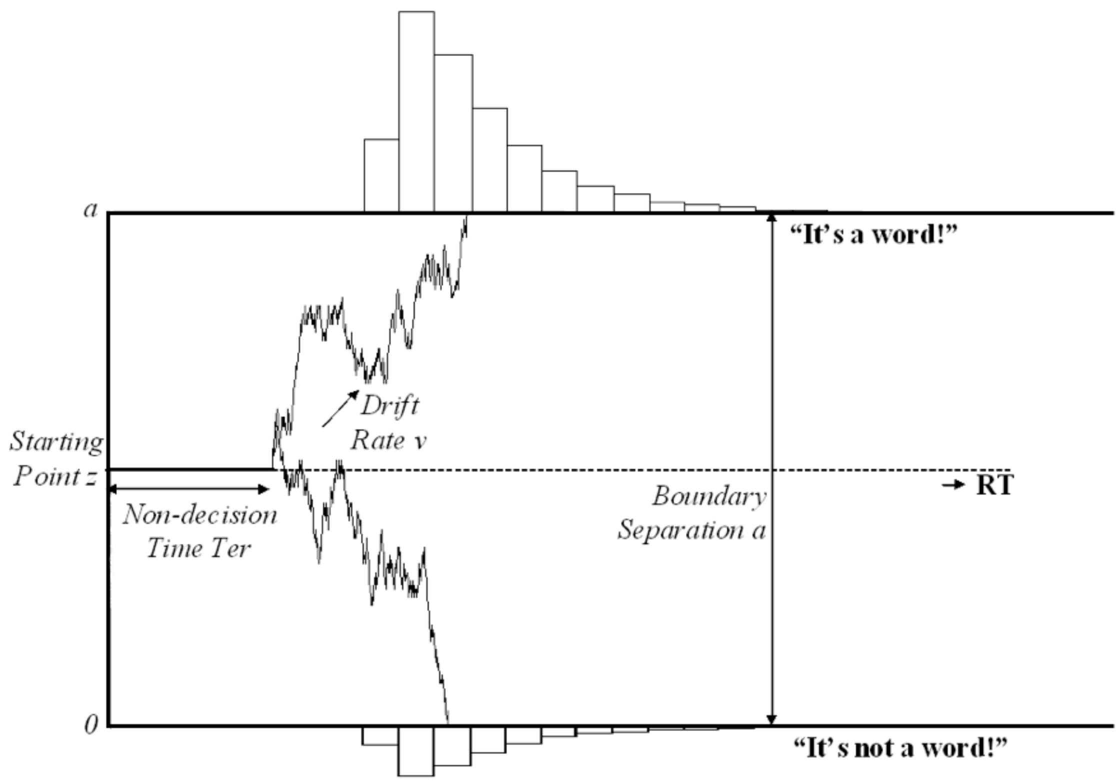

Figure 4. The diffusion model and its key parameters. Evidence accumulation begins at z, proceeds over time guided by drift rate v, is subject to random noise, and stops when either the upper or the lower boundary is reached. The distance between the boundaries is a. The predicted RT is just the accumulation time, plus a constant value for non–decision processes T_{er}.

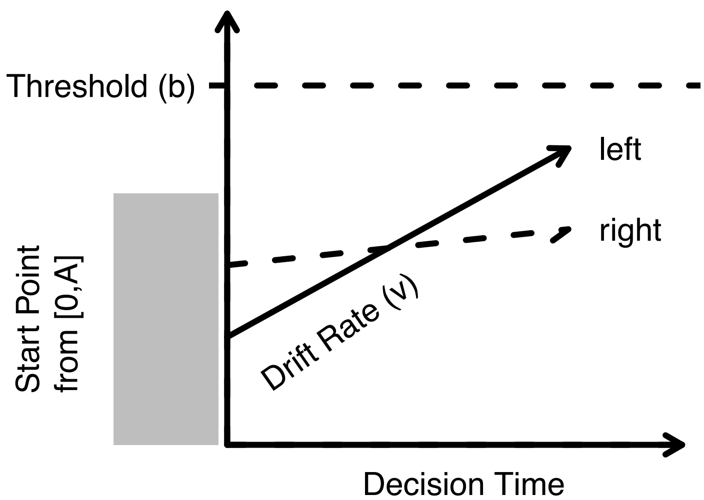

Figure 5. A typical LBA decision. In the illustrated trial, evidence is gathering more quickly in favor of deciding that "left" than "right". The decision will be made as soon as an accumulator reaches the threshold, shown by the dashed line.